Evaluation of two pipelines for calling Copy Number Variants (CNVs) in whole exome data from a cohort of Portuguese azoospermic men

Arti Bandhana,
Masters in Mathematical Engineering
Department of Mathematics
2017

Orientador
Dr Alexandra Lopes
Researcher
The Institute of Molecular Pathology and Immunology of the University of Porto (Ipatimup)

Coorientador
Dr Eduardo Sousa
Post-Doc Researcher
Centre of Molecular and Enviromental Biology

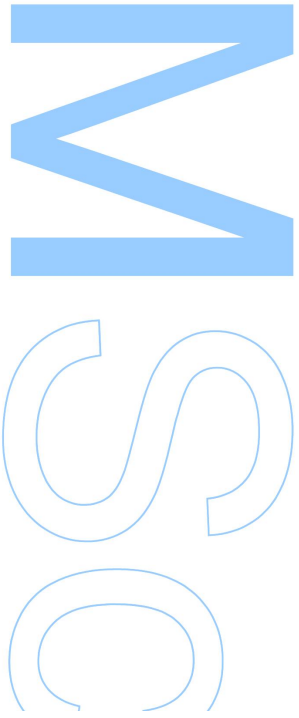
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



UNIVERSITY OF PORTO

Evaluation of two pipelines for calling Copy Number Variants (CNVs) in whole exome data from a cohort of Portuguese azoospermic men

by

Arti Bandhana

Supervisor : Dr Alexandra Lopes

Co-supervisor : Dr Eduardo Sousa

A thesis submitted in partial fulfillment for the degree of Masters in Mathematical Engineering

in the

Faculty of Sciences
Department of Mathematics

July 2017

Abstract

Male infertility is a disorder affecting a large number of men of reproductive age and it often manifests through low sperm count. Azoospermia is its most severe form, characterized by nil sperm counts and it can be caused by chromosome defects (aneuploidy) and microdeletions on the Y-chromosome. However in most cases the underlying cause remains unknown. In order to characterize the unknown genetic architecture of azoospermia we aim to detect rare deletions and duplications (CNVs) in the coding regions of the genome (exome) in a large cohort of azoospermic men. In a former study performed in Portuguese and American infertile men we have detected an excess of structural genetic variants (duplications and deletions) across the whole genome. The current project which is in underway aims at sequencing the exome of 1000 infertile men from Europe and USA, as part of an international consortium - GEMINI (Genetics of Male Infertility Initiative), funded by the NIH (National Institutes of Health, USA). To do this we rely on recently developed algorithms for statistical analysis and modelling of a large amount of exome data. In this study we explore two such algorithms; XHMM and CoNIFER to detect the deletions and duplications. Both of of the algorithms are based on the read depth calculation on the exome sequencing data. XHMM uses principal component analysis to detect data variation and then normalizes it while CoNIFER uses singular value decomposition to do the same. We ran the two callers on a dataset of 94 samples. XHMM detected a total of 3185 variants with 1527 deletions and 1658 duplications. CoNIFER detected a total of 677 variants with 98 deletions and 579 duplications. Our results show that the detections made by the two callers have a wide variation and further improvements are needed to the decrease the variation gap which would result in robust calling of the variants.

Resumo

A infertilidade masculina afeta um grande número de homens em idade reprodutiva e manifesta-se muitas vezes através de uma baixa contagem de espermatozóides no ejaculado. Azoospermia é a forma mais grave de infertilidade masculina, caracterizada por contagens de espermatozóides nulas e pode ser causada por defeitos cromossómicos (aneuploidia) e microdeleções no cromossoma Y. No entanto, na maioria dos casos a causa subjacente permanece desconhecida. Para contribuir para a caraterização da arquitetura genética da azoospermia, temos como objectivo detectar deleções e duplicações raras (CNVs) nas regiões codificantes do genoma (exoma) numa grande coorte de homens azoospérmicos. Num estudo anterior realizado em homens inférteis portugueses e americanos, detetamos um excesso de variantes genéticas estruturais (duplicações e deleções) em todo o genoma. O projeto que agora estamos a desenvolver visa a sequenciação do exoma de 1000 homens inférteis da Europa e dos EUA, como parte de um consórcio internacional - GEMINI (Genetics of Male Infertility Initiative), financiado pelo NIH (National Institutes of Health, EUA). Para o conseguir utilizamos algoritmos desenvolvidos recentemente para análise e modelação estatística de uma grande quantidade de dados sequenciação de exomas. Neste estudo, exploramos dois desses algoritmos: XHMM e CoNIFER para detectar as deleções e duplicações. Ambos os algoritmos são baseados no cálculo de profundidade de leitura nos dados de sequenciação de exoma. O XHMM usa a análise de componentes principais para detectar a variação dos dados e depois faz uma normalização, enquanto o CoNIFER usa a decomposição de valores singulares para o mesmo efeito. Nós analisamos os dados de sequenciação de 94 indivíduos com os dois programas. O XHMM detectou um total de 3185 variantes, sendo1527 deleções e 1658 duplicações. O CoNIFER detectou um total de 677 variantes, sendo 95 deleções e 579 duplicações. Os nossos resultados mostram que a deteção feita pelos dois programas tem uma ampla variação e são necessárias melhorias para diminuir esta variação e obter uma deteção mais robusta destes variantes.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr Alexandra Lopes and co-supervisor Dr Eduardo Sousa for their useful comments, remarks and engagement through the learning process of this master thesis. Their expert opinion and support always steered me in the right direction.

Contents

List of Figures

List of Tables

Abbreviations

AZF	**AZ**oospermia **F**actor
BAM	**B**inary **A**lignment **M**ap
CNV	**C**opy **N**umber **V**ariation
CoNIFER	**Co**py **N**umber **I**nference **F**rom **E**xome **R**eads
DNA	**D**eoxyribo**N**ucleic **A**cid
FTPS	**F**ile **T**ransfer **P**rotocol **S**ecure
GATK	**G**enome **A**nalysis **T**oolkit
HTCF	**H**igh **T**hroughput **C**omputing **F**acility
NGS	**N**ext **G**eneration **S**equencing
PCA	**P**rincipal **C**omponent **A**nalysis
RPKM	**R**eads **P**er **K**ilobase **M**illion
SNV	**S**ingle **N**ucleotide **V**ariant
SNP	**S**ingle **N**ucleotide **P**olymorphism
SVD	**S**ingular **V**alue **D**ecomposition
VCF	**V**ariant **C**alling **F**ile
XHMM	e**X**ome - **H**idden **M**arkov **M**odel

Chapter 1

Introduction

Human beings are different from each other and much of these differences have a genetic basis; differences in phenotype caused by differences in genotype. While some differences can be readily observed in the family such as hair, eye, stature and skin color others are concealed and a subject of research. Phenotypic traits of humans are controlled by a combination of inherited and environmental factors, stochastic developmental events and molecular processes. The easiest traits to dissect genetically are those determined in large part by a single gene, called the Mendelian traits. For example, some genetic differences are directly responsible for diseases such as Huntington disease and cystic fibrosis, which can be passed to the offspring and the chances of it affecting the child depends on whether the particular allele is dominant or recessive. An individual with one dominant and one recessive allele for a gene will have the dominant phenotype. They are generally considered carriers of the recessive allele: the recessive allele is there, but the recessive phenotype is not[1]. However many of the phenotypic traits which hold the most interest to researchers are complex, determined by multiple genes and the environment. Understanding human phenotypes and associated diseases also depends on the knowledge that humans shared a common ancestor at least 200 ky ago and with every other species on the planet several million years ago. In order to be able to identify genetic variants associated with a disorder it is also necessary to thoroughly characterize the human genome and the genetic diversity in human populations. A comparative analysis with evolutionarily close model organisms is also very informative

[1] http://learn.genetics.utah.edu/content/basics/patterns/

to identify conserved regions that have not been characterized but may perform an important function in both species.[1]

1.1 Human genome and structure

A genome is an entire set of hereditary instructions for building, running, and maintaining an organism, and passing life on to the next generation, in other words the complete genetic material of a living being. In humans and most organisms with the exception of some viruses, the genetic material is made up of deoxyribonucleic acid (DNA). The genome contains genes, which are packaged in chromosomes and affect specific characteristics of the organism [2]. This relationship can be represented as a set of Chinese boxes nested inside one another as shown in fig 1.1. The largest box represents the genome, which is divided into chromosomes, chromosomes contain genes, and genes are made of DNA.

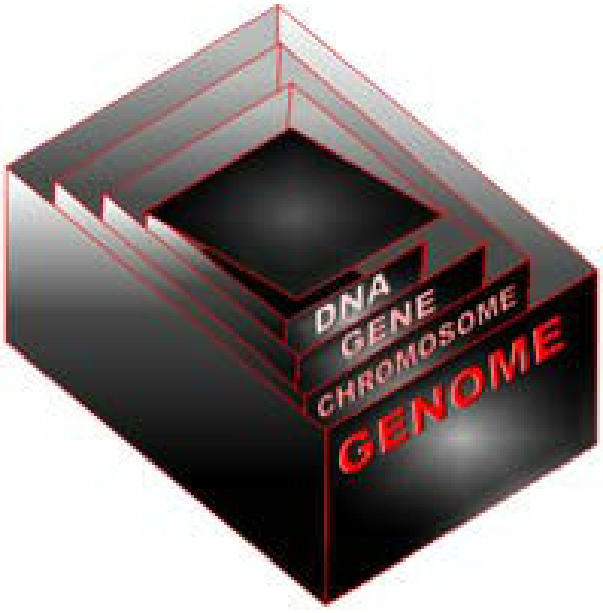


FIGURE 1.1: Genome News Network

DNA is an extraordinary informational macromolecule that carries instructions to the cells and provides a means for this set of instructions to be passed to the daughter cells when a cell divides. DNA is a polymer and its monomeric subunits are molecules called nucleotides, which are of four varieties and differ in portions known as bases. The bases as shown in fig 1.2 are adenine (A), guanine (G), cytosine(C) and thymine (T) and it is

[2] http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp1_1_2.shtml

the sequence of these pairs of nucleotide molecules that carries the genetic information. Each of these bases is joined to a sugar molecule, deoxyribose and each deoxyribose has a phosphate group attached to it. The sugar and phosphate group do not carry any genetic information and only play a structural role.

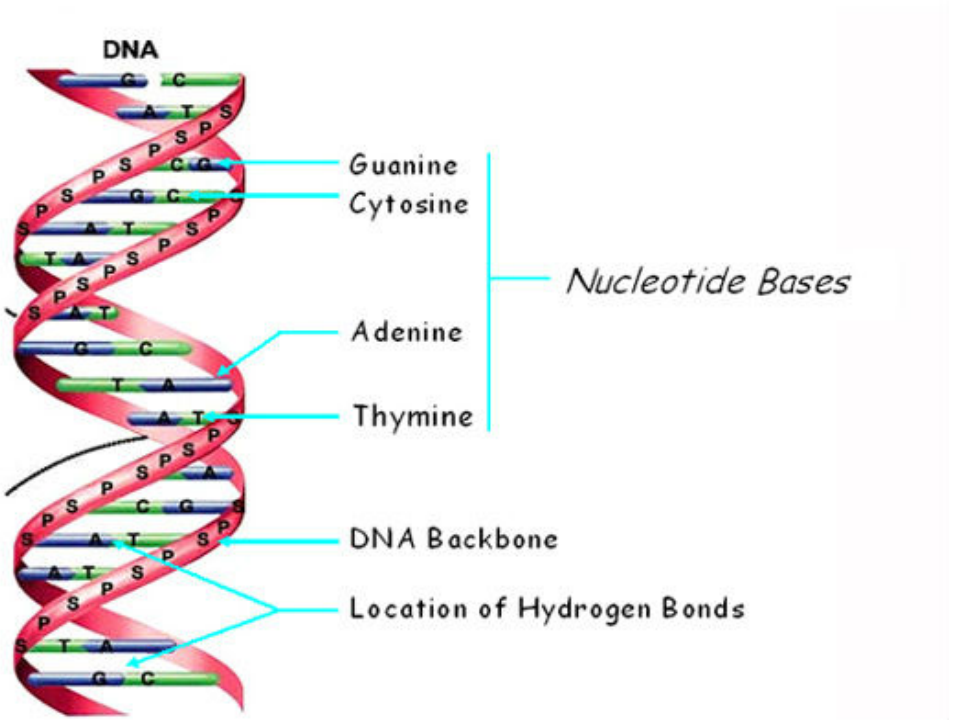


FIGURE 1.2: How The First DNA Evolved Vedas

Each human being contains a slightly different version of the human genome, but the differences between two people are much smaller than the genome differences between humans and their closest species i.e the chimpanzee[1].

The genetic material is stored in two organelles: nucleus and mitochondria. The nuclear genome consists of 3.2 million base pairs (bp) which are packed in 22 pairs of autosomes and two sex chromosomes, X and Y. Human chromosomes are not of equal sizes; the smallest, chromosome 21, is 54 million bp long; the largest, chromosome 1, is almost five times bigger with 249 million bp[3].

[3]http://www.actabp.pl/pdf/3_2001/587-598s.pdf

1.2 Sex Chromosomes and male infertility

Human beings like most of the animals are diploid, i.e we have two copies of the genome in the somatic cells, one from each parent. There are however two segments of DNA that are inherited from each parent differently, and determine the sex of the offspring, the X and Y chromosomes. Fig 1.3 shows the 22 chromosomes with the sex chromosomes highlighted in the lower right.

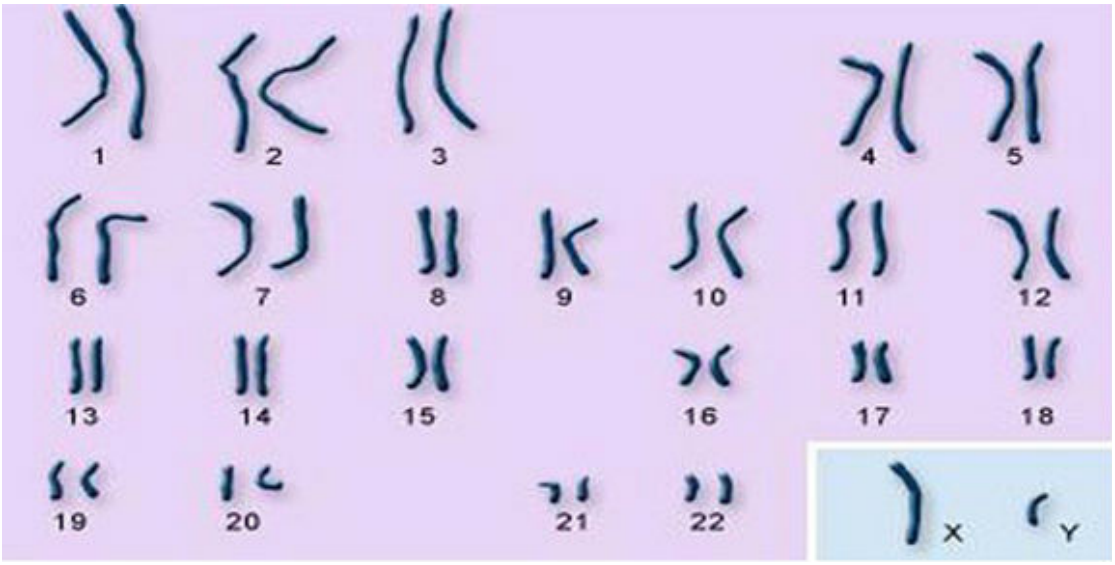


FIGURE 1.3: Autosomes and the sex chromosomes; Numbers 1 to 22 represent somatic diploid chromosomes while X and Y represent germ-line haploid chromosome, Credit: U.S. National Library of Medicine

While the X chromosome is present in both parents, the Y chromosome is haploid, i.e one copy of the genome which is malespecific and sex-defining. The Y chromosome spans more than 59 million building blocks of DNA (base pairs) and represents almost 2 percent of the total DNA in cells. Each person normally has one pair of sex chromosomes in each cell. The Y chromosome is present in males, who have one X and one Y chromosome, while females have two X chromosomes. The Y chromosome contains 50 to 60 genes that provide instructions for making proteins. Because only males have a Y chromosome, the genes on this chromosome are mostly involved in male sex determination and development. The SRY gene determines sex, thus being responsible for the development of a fetus into a male. Other genes on the Y chromosome are important for male fertility and the microdeletions on the AZF (Azoospermia Factor) regions are a well-established cause of male infertility causing severe oligozoospermia (sperm counts

lower than 5 million spermatozoa per mL) or azoospermia (absence of sperm in the ejaculate)[2].

About 15% of the couples globally (amounting to approximately 48.5 million couples) are affected by infertility and out of these 20-30% of the cases occur due to male infertility alone. On a global scale, accurate information regarding rates of male infertility is acutely lacking, however based on the amalgamation of numbers it is estimated that about 30 million men suffer from infertility worldwide with the highest rate in Africa and Easten Europe.[3]

More than 90% of male infertility cases are due to low sperm counts, poor sperm quality, or both. After excluding anatomical problems, hormonal imbalances, and known genetic defects the majority of cases remain as idiopathic, i.e. with an unknown cause. In these cases a genetic defect may underlie the low sperm counts or even the absence of sperm in the ejaculate-azoospermia. In spite of several decades of research on the genetics of male infertility only a few convincing candidate genes have been found and more recently genome-wide studies have provided evidence for a heterogeneous genetic etiology with likely many genomic regions contributing to the disease[[4], [5], [6]]. In these studies the sex chromosomes stand out as harboring many variants associated with low to nil sperm counts.

In this work the whole genome will be studied with an emphasis on the sex chromosomes and in particular the Y chromosome, which harbors male candidate genes for male infertility.

1.3 Genetic variants and disease

As previously stated humans are genetically diverse and in any population different types of DNA variants can be found such as:

- SNP/SNV - a single-nucleotide polymorphism (SNP) or single nucleotide variant (SNV) involves a change in a single base
- Insertion-deletion (indel) - a type of genetic variation where a specific nucleotide sequence is either present (insertion) or absent (deletion).

- Copy number variant(CNV) - a type of structural variant generated through deletions or duplications in the number of copies of specific regions of DNA [7]. It comprises of insertions, deletions and duplications of DNA segments of minimum size 50bp and may span up to 1Mbp, collectively spanning about 12% of human genome[8].

Genetic variants range from rare (<1%) to frequent (>5%) in a given population and their frequency may vary across different human populations. Common variants are most often non deleterious, even though it is believed that a set of common variants may underlie susceptibility to common diseases. Rare variants may simply be younger in the population or may be kept at low frequencies because they are deleterious, interfering with gene function and reducing the fitness or fertility of the carrier. While rare deleterious variants with a strong effect have been mostly associated with rare diseases it is also accepted that a collection of rare deleterious variants can also contribute to common diseases[9].

1.4 Why study CNVs?

During the past several years, thousands of new variations in repetitive regions of DNA have been identified, leading researchers to believe that copy number variations are a component of genomic diversity as important as single nucleotide polymorphisms (SNPs)[4]. Summarizing the previous sections, the human genome is comprised of 6 billion chemical bases (or nucleotides) of DNA packaged into two sets of 23 chromosomes, one set inherited from each parent. The DNA encodes 30,000 genes (fig 1.4). Over the years it was believed that these genes were always present in two copies however recent studies have revealed that large segments of DNA ranging in size from thousands to millions of DNA bases vary in copy- number. Due to its encompassing attribute CNVs play an important role in both human disease and drug response. The human CVN map is being used to exclude variation found in unaffected individuals of certain disease or disorders helping the researchers to target the regions that might be involved[5]. In addition to that the data generated can help to make a more accurate and complete human genome reference.

[4]https://www.nature.com/scitable/topicpage/copy-number-variation-445
[5]http://cnv.gene-quantification.info/

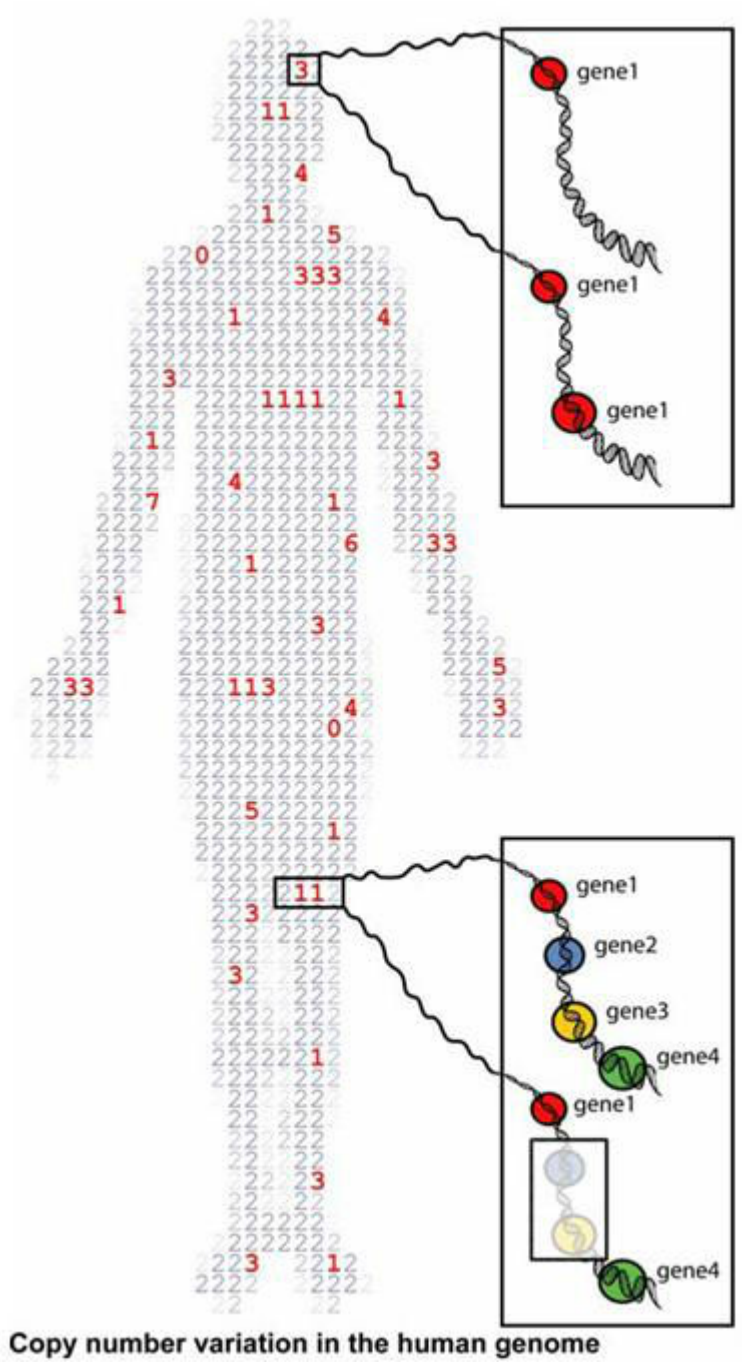


FIGURE 1.4: The 30,000 genes are usually present in two copies. New studies have unveiled a new a map of the genome by cataloguing DNA and genes variable in copy number (numbers other than 2 highlighted in red) across world wide populations. Duplication of the gene(top) and deletion of two genes(bottom) are depicted, Credit: Gene Quantification

There have been approximately 2000 CNVs detected till date with the average size being 250,000 bases[6]. Although most CNVs are benign variants and have no known association with any disorder, some interfere with the function of the encompassed genes or of

[6][7]

the genes in their vicinity and cause disease. Compared to other genetic variants, CNVs are larger in size and can often involve complex repetitive DNA sequences. Traditionally, larger CNVs were identified using cytogenetic technologies such as karyotyping and fluorescence in situ hybridization (FISH), and later more accurate array-based comparative genomic hybridization (aCGH) and single-nucleotide polymorphism (SNP) array approaches have been applied. However, these approaches suffer from several inherent drawbacks, viz., hybridization noise, limited genome coverage, low resolution, and difficulty in detecting novel and rare CNVs[10]. With the advent of next generation sequencing (NGS) this approach has become the method of choice for genome-wide structural variation analyses of rare and unknown variants. NGS approach offers advantages such as higher coverage and resolution, accurate detection of copy number and breakpoints of CNVs of varying lengths and the ability to identify novel CNVs.

Compared to the entire genome sequencing, whole exome sequencing is less costly and more time effective. Exons are the coding portion of the genome and represent less than 2% of the genome, but contain 85% of known disease-related variants making whole-exome sequencing a cost-effective alternative to screen CNVs associated with disease in a large number of samples. It produces a smaller, more manageable data set for faster, easier analysis compared to whole-genome approaches and hence it is becoming increasingly popular. Despite of these advantages, detecting copy number variants from exome sequencing is challenging because of the noncontiguous nature of the captured exons and various algorithms have been formulated to address these challenges.[10]

1.5 Tools for detecting CNVs

The NGS based CNV detection methods can be categorized into five different strategies, including: paired-end mapping (PEM), split read (SR), read depth (RD), de novo assembly of a genome (AS), and combination of the above approaches (CB) (Fig 1.5). Indeed, different strategies have their own advantages and limitations. Though there has been great progress in each category, none of the methods could comprehensively detect all types of CNVs[10].

A. Paired-end mapping (PEM) strategy detects CNVs through discordantly mapped reads. A discordant mapping is produced if the distance between two ends of a read

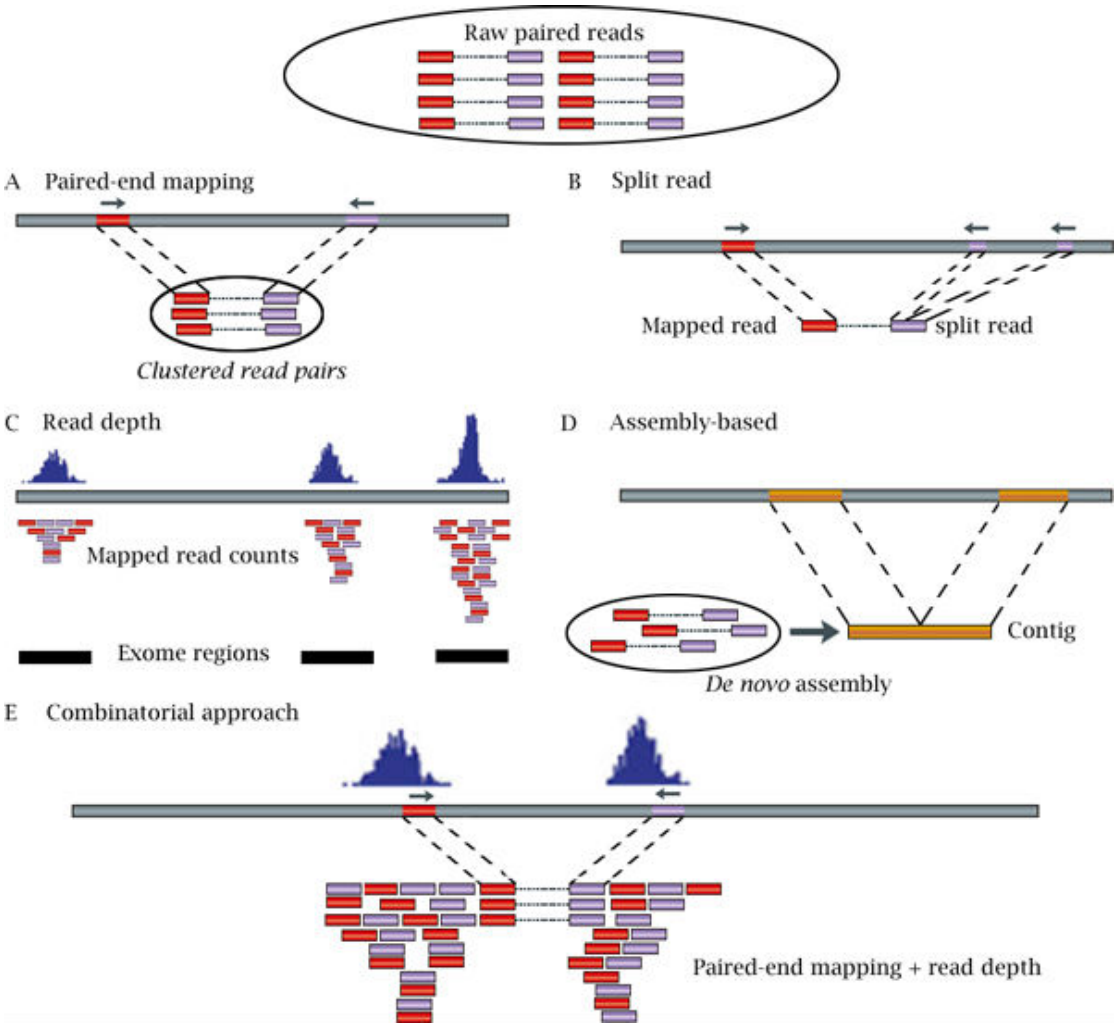


FIGURE 1.5: Five approaches to detect CNVs from NGS short reads

pair is significantly different from the average insert size. B. Split read (SR)-based methods use incompletely mapped read from each read pair to identify small CNVs. C. Read depth (RD)-based approach detects CNV by counting the number of reads mapped to each genomic region. In the figure, reads are mapped to three exome regions. D. Assembly (AS)-based approach detects CNVs by mapping contigs to the reference genome. E. Combinatorial approach combines RD and PEM information to detect CNVs[10].

1.6 Scope

This research/thesis aims to explore different software tools, which are being utilized by researchers for the detection of CVNs from whole exome sequencing (WES) data. The

goal here is to test different pipelines for CNV detection, analyze and compare the results and assist in the choice of software most fit for performing association studies with rare CNV variants in severe cases of male infertility without a known cause (idiopathic).

Since our data is for exome sequence, therefore we used software tools that are based on read-depth approach. This approach follows a four-step procedure to discover CNVs: mapping, normalization, estimation of copy number, and segmentation[10]. The sequencing reads were provided in .bam format (Binary Alignment Map) therefore we restricted our choice of tools to read-depth approach which takes bam files as input. The two tools that were tested are : XHMM (eXome-Hidden Markov Model) and CoNIFER (Copy Number Inference From Exome Reads)

Chapter 2

Methodology

In order to satisfy the goals/objectives outlined above an inductive approach was applied.The samples were sequenced on a Illumina HiSeq 4000. The capture reagent was an in-solution custom exome capture reagent designed at Washington University, targeting approximatelly 19 000 genes from the generating paired-end libraries. A pilot study was conducted in collaboration with the Washington University in St Louis (MO, USA) on 94 (93, and one reference sample NA12878), Portuguese azoospermic patients. In the scope of this study another 400 samples will be sequenced and the pipeline chosen after the pilot study will be then applied to all the samples.

2.1 Software resources

- HTCF computer cluster (https://htcf.wustl.edu/docs/) - Due to the large size of the files, the variety of software resources needed to handle the data and the need for many jobs to be run simultaneously the analysis had to be conducted on a high performance computer cluster based and maintained by the Washington University. Most of the required software is pre-loaded on the HTCF cluster. It was straightforward to load the required modules for each analysis. Server access was granted for the remote connection and FTPS client software was used to get access to the files.

- XHMM[11] (http://atgu.mgh.harvard.edu/xhmm/tutorial.shtml) - The first pipeline tested was developed at the Broad Institute and is based on a statistical tool

(exome hidden Markov Model) that uses principal component analysis (PCA) to normalize exome read depth and a Hidden Markov Model (HMM) to discover exon-resolution CNV and genotype variation across samples. The software had to be downloaded from the XHMM website and installed on the server.

- CoNIFER[12] (http://conifer.sourceforge.net/tutorial.html) - For the second pipeline a python based script was used which used singular value decomposition to eliminate the biases in the exome data and make calls.
- Genome Analysis Toolkit (GATK) (https://software.broadinstitute.org/gatk/) - For the calculation of depth of coverage, one of GATK modules was used.
- Bedtools (http://bedtools.readthedocs.io/en/latest/) - Bedtools allows to find overlapping regions between datasets and perform statistical tests to determine if certain classes of genomic regions (such as genes, repetitive regions, etc) are enriched in a dataset.
- R - https://www.r-project.org/
- University of California Santa Cruz (UCSC) Genome Browser (https://genome.ucsc.edu/)
- Ensembl (http://www.ensembl.org/index.html)
- PLINK (http://zzz.bwh.harvard.edu/plink/)

2.2 Description of the file formats

- The main output file for XHMM is PT.DATA.xcnv, which contains one line for each CNV called in an individual. The columns in this file (table 2.1) denote the quantities for that CNV which will be used as a reference throughout this thesis.

- PT.DATA.vcf - another output file by XHMM. The file file contains haploid genotypes for each individual, or a missing genotype if the normalized read depth is not definitive.
- calls.txt - the main output file for CoNIFER which contains the sample ID, chromosome name & coordinate and variant type.

Table 2.1: Xhmm output explanation

SAMPLE	sample name in which CNV was called
CNV	type of copy number variation (DEL or DUP)
INTERVAL	genomic range of the called CNV
KB	length in kilobases of called CNV
CHR	chromosome name on which CNV falls
MID_BP	the midpoint of the CNV (to have one genomic number for plotting a single point, if desired)
TARGETS	the range of the target indices over which the CNV is called
NUM_TARG	# of exome targets of the CNV
Q_EXACT	Phred-scaled quality of the exact CNV event along the entire interval
Q_SOME	Phred-scaled quality of some CNV event in the interval
Q_NON_DIPLOID	Phred-scaled quality of not being diploid, i.e., DEL or DUP event in the interval
Q_START	Phred-scaled quality of left breakpoint of CNV
Q_STOP	Phred-scaled quality of right breakpoint of CNV
MEAN_RD	Mean normalized read depth (z-score) over interval
MEAN_ORIG_RD	Mean read depth (# of reads) over interval

- BAM - the samples were provided in .bam format i.e a binary format for storing sequence data.
- VCF - a text file format used for storing gene sequence variations. It contains meta-information lines, a header line, and data lines each containing information about a position in the genome.
- probes.txt - it is a text file that contains chromosome coordinates (chr:start-stop coordinates) for the targets in the exome capture.

Chapter 3

Algorithms used

3.1 XHMM (eXome-Hidden Markov Model)

Copy-number variants (CNVs) have surfaced in the past decade as a category of structural genetic variants that play a key role in human health and common disease. Numerous tools exist for the detection of CNVs, however they are influenced by a greater margin of error than that associated with the calling of other types of genetic variants, which makes it difficult for the researchers to analyze the results and formulate a sound conclusion. To overcome this obstacle, a new statistical tool has been developed XHMM which uses principal component analysis (PCA) to normalize exome read depth and a hidden Markov model (HMM) to discover exon-resolution CNV and genotype variation across samples [11]. We applied XHMM to detect deletions and duplications in a sample of 93 Portuguese azoospermic patients. One experimental control sample was included in the analysis.

3.1.1 XHMM workflow

The flow chart in fig 3.1 outlines the main steps of the XHMM pipeline

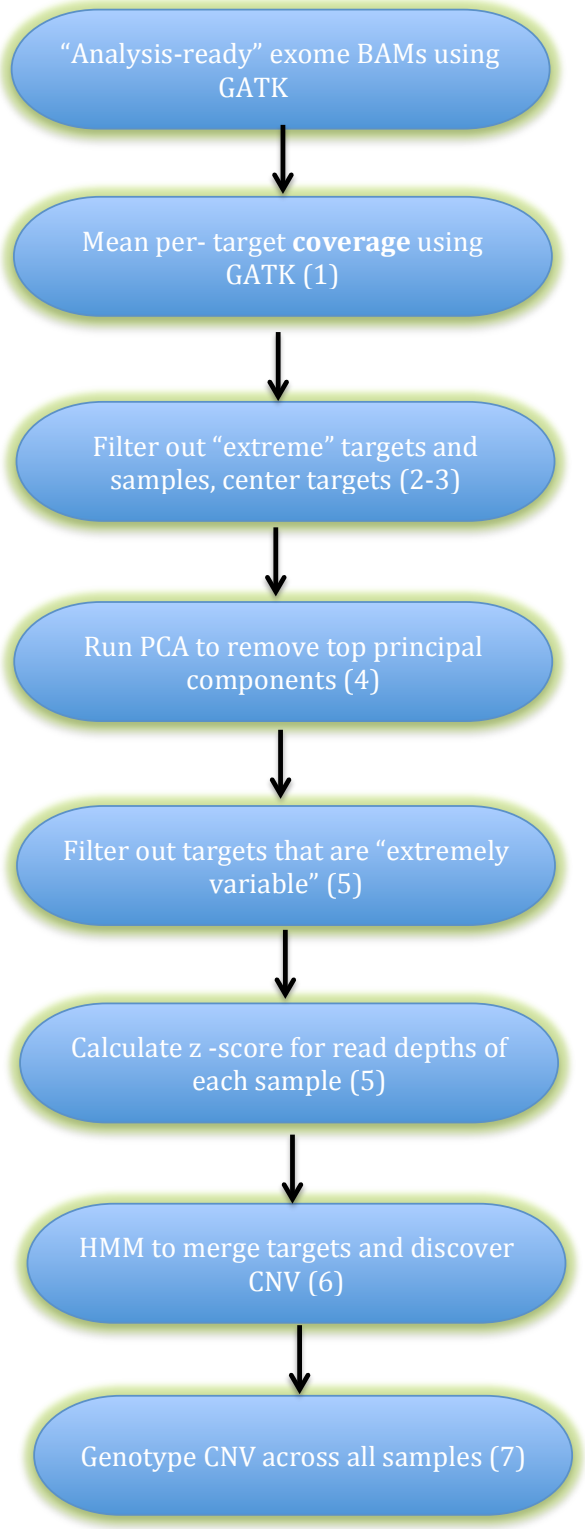


FIGURE 3.1: Xhmm work flowchart

3.1.2 Overview of the steps

A total of 94 samples were analyzed which were already aligned to the reference genome(HGRC38 assembly) and in BAM format. BAM files of "analysis-ready reads" used had been previously generated using the "best practice" combining bwa, Picard and GATK pipelines.

The pipeline consists of the following stages:

- The first step of the workflow is to run GATK (Genome Analysis Tool Kit) to obtain the depth of coverage for each target for the 94 samples. Due to the very large size of the BAM files, they had to be split into ten groups to allow running the analysis in parallel in the cluster. Subsequently this step combines the Depth of Coverage outputs. It extracts the mean coverage for each exon interval for each sample and merges them all into a single file.

- The next two steps are optional, namely i) calculate the GC content of the targets with GATK and ii) calculate sequence complexity of targets with Plink/Seq. These steps aim at identifying those targets that may be more prone to biases in the following analyses due to their biological properties. The output file from the former would be used in the filtering and normalization step. It calculates per-target GC content and creates a list of targets with extreme GC content. The latter calculates the fraction of repeat masked bases in each target and creates a list of targets with low complexity (high percentage of repeated bases) which is used in filtering. Given that the proposed Plink/Seq procedure generated several errors that have been also reported by other groups and that these particular properties of the targets can be verified after the CNV calling we did not include these steps in our pipeline.

- This stage removes samples and targets with outlier read depth values and then mean-centers the targets in preparation for the PCA normalization in the succeeding step. This step allows to exclude targets based on sequence complexity and GC content. We filtered only the targets with extreme GC content.

- The next step is to run a principal component analysis (PCA) on the mean-centered data to determine the strongest independent components contributing to the variance observed in the data. In any genome-wide analysis only a small proportion of

the regions under analysis is expected to vary significantly between individuals and most of the variance between samples will be experimentally driven and should be removed from the analysis. Thus, once the main sources of variance in read depth in our dataset were identified by PCA they were removed.

- The targets that had a very high variance after PCA normalization were considered outliers and were also removed. Here we used XHMM to calculate the z-scores of the read depths for each sample by centering relative to all target read depths. On a first approach we only kept for the next steps of the analysis the targets that had a post-normalization standard deviation value greater than 30, as recommended in the default settings.

From the pre-normalized read depths, we removed the same targets and samples that were removed during the normalization process. The original read depth data was filtered to restrict to the same sample and targets as the normalized data.

- After read depth normalization, the last step is to make the CNV call for each sample. This is the step where hidden Markov model is used. The main output file gives the details of the sample, type of CNV (DEL or DUP), interval, chromosomes, KB and other details that are necessary for the downstream analysis.
- It is beneficial to collect all variation called across multiple samples and then uniformly regenotype these variants in each sample. This step runs the HMM algorithm to quantitatively genotype each called CNV in all samples. Since our major focus was on the number of these variants called and not their genotype, this result was not investigated in detail. It was only used for intersection of files for next stages.

3.1.3 Quality Control of the Analysis

Several aspects are crucial for the quality control of the data generated by whole exome sequencing.

❧ Mean sample and mean target coverage:

Sample coverage is a crucial factor that allows the evaluation of sample quality and mean target coverage evaluates homogeneity of the exome capture experiment. Both allow predicting the success of the CNV calling. In fact, deletions and duplications will be called in regions with significantly lower or higher read depth compared to other samples and the higher is the overall sample/target coverage the more robust will the CNV calling be. The plots in fig 3.2 and 3.3 depict the mean sample coverage and mean target coverage in our pilot experiment.

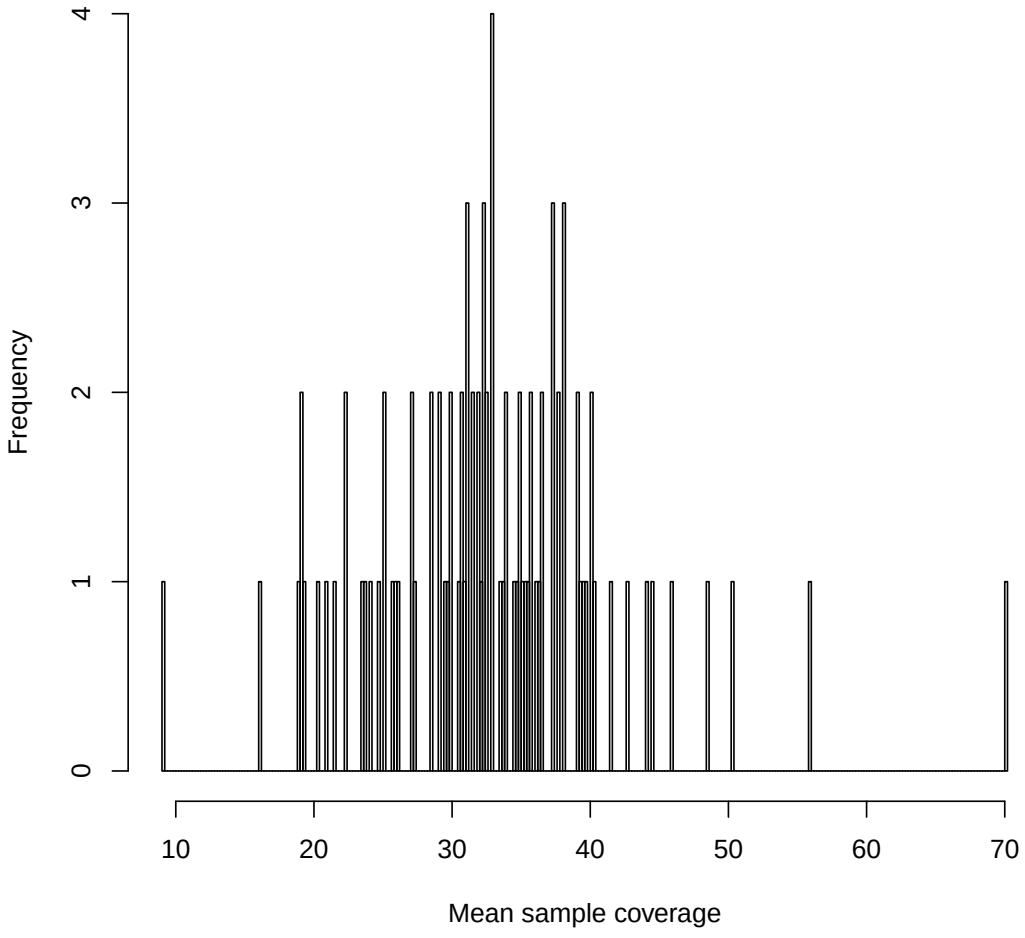


FIGURE 3.2: Mean sample coverage for the 94 samples with respect to the frequency. This shows the sample distribution of exome-wide sequencing coverage, where each per-sample coverage value is the mean of the coverage values calculated for each exome target. In this experiment, we sequenced each sample to a mean coverage of 30, so that we expect a typical sample to indeed have 30 reads covering an average base in an average exome target. Here we can see that the read depth is normally distributed with the mean around 30 and a few outlier samples with much higher and much lower mean coverage, as expected

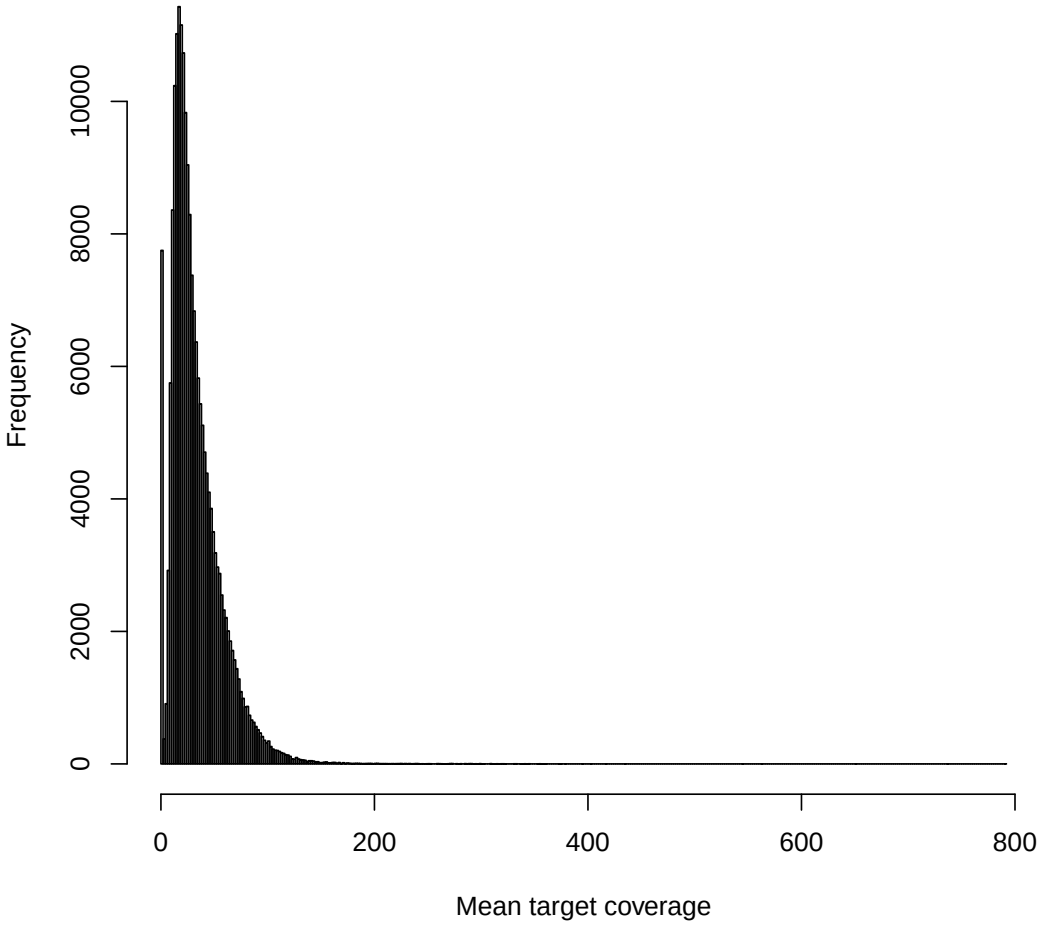


FIGURE 3.3: Mean target coverage for the 94 samples with respect to the frequency. This plot gives the target-wide distribution of coverage (over all samples). That is, each per-target coverage value is the mean of the per-sample coverage values at that target. As above, since our goal was to have 30 coverage exome-wide, we would expect each target to have around 30 coverage, but we see here that there is high variability in target coverage. For example, some targets have approximately 180 coverage (averaged over all samples), and we also see a non-trivial number of targets that have 0 coverage for all samples (e.g., targets where capture has presumably failed). This is expected as the efficiency of target capture is not equal for all targets.

❧ Principal Component Analysis for data Normalization:

The XHMM pipeline implements sample normalization using principal component analysis and removing from the analysis the k principal components that contribute most to the variance observed between samples (fig 3.4).

Each Principal Component is correlated with a feature that is variable in the samples

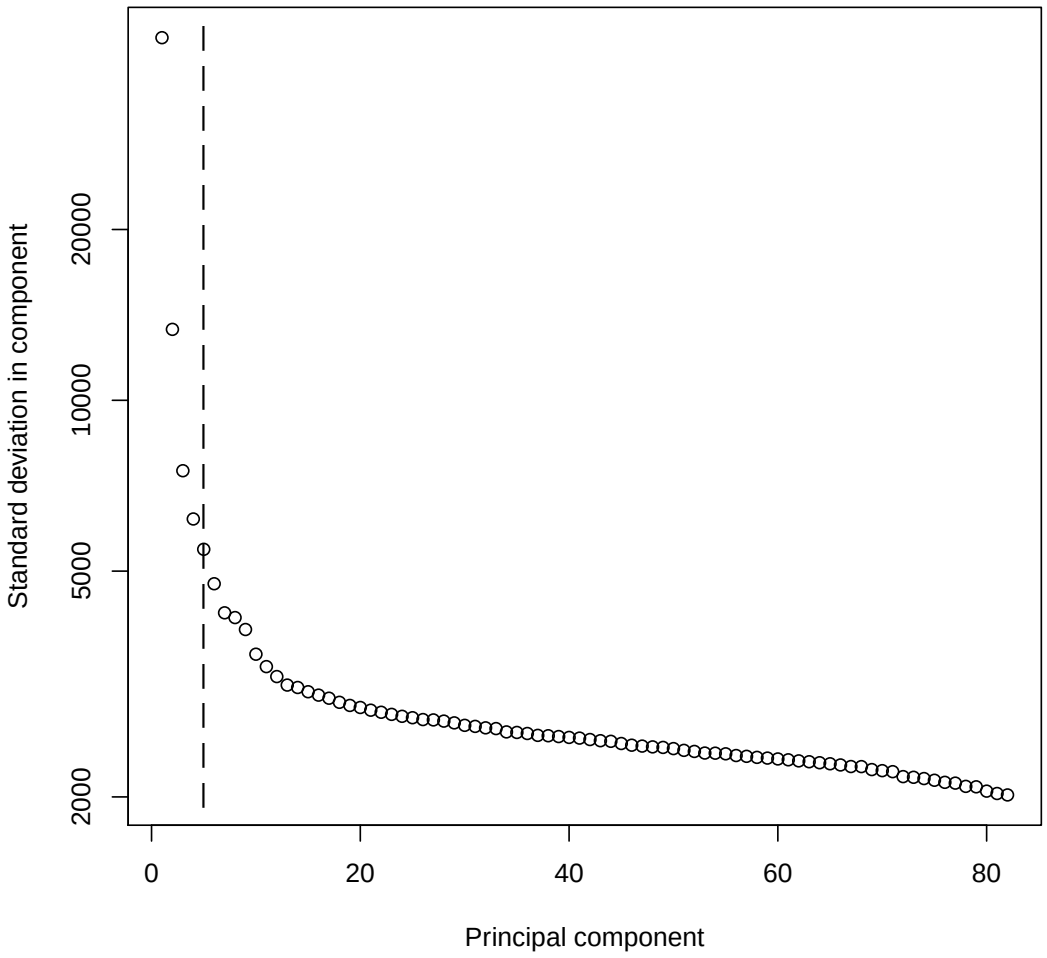


FIGURE 3.4: Scree plot for the PCA. This plot shows the standard deviation of the read depth independently ascribed to each of the principal components. This case is typical, where we see that the cut-off automatically detected by XHMM corresponds to a significant drop in the variance (an elbow in the curve).

or in the targets (fig 3.5). In our dataset sample mean read depth is the only feature that contributes to the PCs and thus the normalization is correcting for between sample variation in read depth, most likely driven by heterogeneity in sample quality.

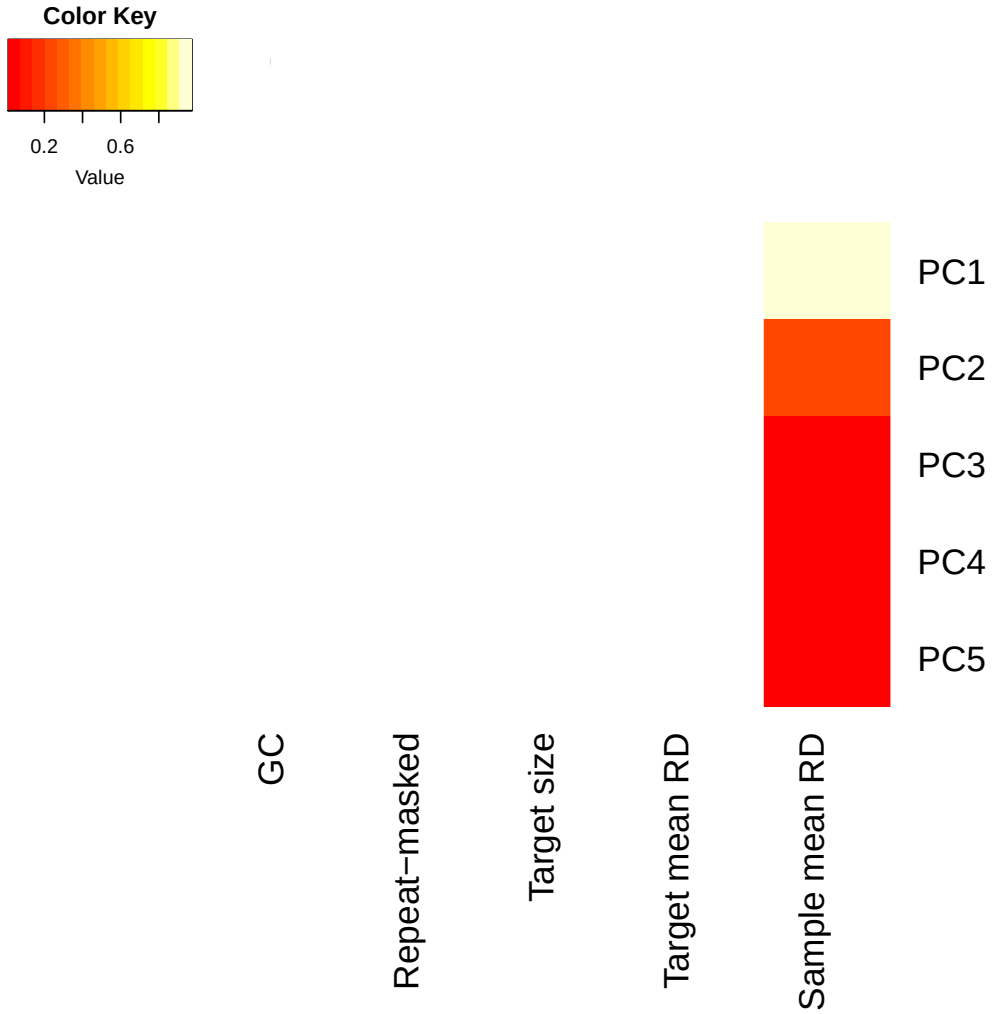


FIGURE 3.5: This plot shows the correlation of sample attributes with each of the five strongest principal components. It can be seen that for this set of data only the sample mean read depth has a correlation with the five strongest PCs and in particular a very high correlation with PC1.

3.2 CoNIFER - (copy number inference from exome reads)

While exome sequencing is becoming increasing popular in the detection of CNVs, some challenges remain due to the sparse and non-uniform nature of the exome capture, which hinder the characterization of the genic copy number variation. To overcome this, a tool was developed which uses singular value decomposition (SVD) normalization to discover rare genic copy number variants as wells as genotype copy number polymorphic (CNP) loci with high sensitivity and specificity from exome sequencing data. CoNIFER uses exome-sequencing data to find CNV and genotype the copy-number of duplicated genes. As exome capture reactions are subject to strong and systematic capture biases between sample batches, singular value decomposition (SVD) was implemented to eliminate these biases in exome data. CoNIFER offers the ability to mix exome sequence from multiple experimental runs by eliminating batch biases.[12]

3.2.1 CoNIFER workflow

The flow chart in fig 3.6 outlines the main steps of the CoNIFER analysis.

3.2.2 Overview of the steps

A total of 94 Bam files already aligned to the reference genome were used to calculate the RPKM (Reads Per Kilobase of transcript per Million mapped reads). Along with the Bam files, a probe file (target definition) was also created which included the chromosome, start & end coordinates of each exon and name of the gene found in that particular region. This was used as a reference while calculating the RPKM and running the main analysis.

The pipeline consists of the following stages:

- A directory was created which would contain the 94 samples and the location of probe file was specified.
- CoNIFER calculates RPKM from aligned and indexed BAM files using the python pysam package(as CoNIFER is entirely a command line python program).

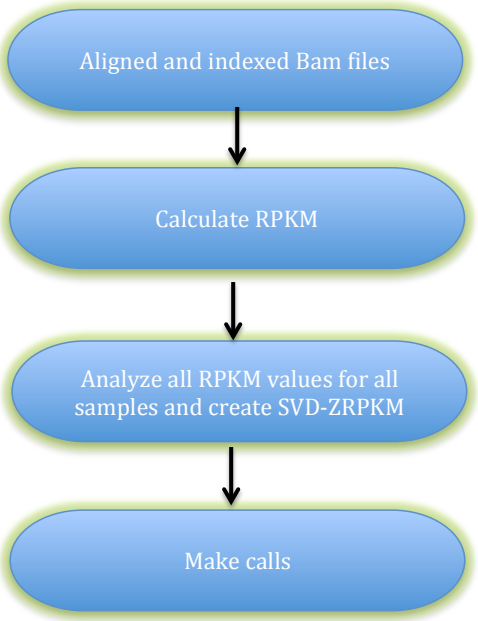


FIGURE 3.6: CoNIFER analysis flowchart

- In this step, the RPKM files are loaded into memory, poorly responding probes are masked, and the Z- and SVD transformations are applied.

- CNV detection was made based on the SVD-ZRPKM: RPKM values were transformed into standardized z-scores (ZRPKM) based on the median and standard deviation across all analyzed exomes and organized into an exon by sample matrix. Biological variation can be seen in the form of rare CNVs as well as common CNPs that is a minor contributor to the overall variance of the exon by sample matrix. This algorithm was formulated to eliminate the strongest variance components. The number of components for elimination is selected based on the scree plot. The strongest k components are removed from the singular values and calls are made. The final values are termed as SVD-ZRPKM values each of which represents the normalized relative copy number of an exon in a sample[12].

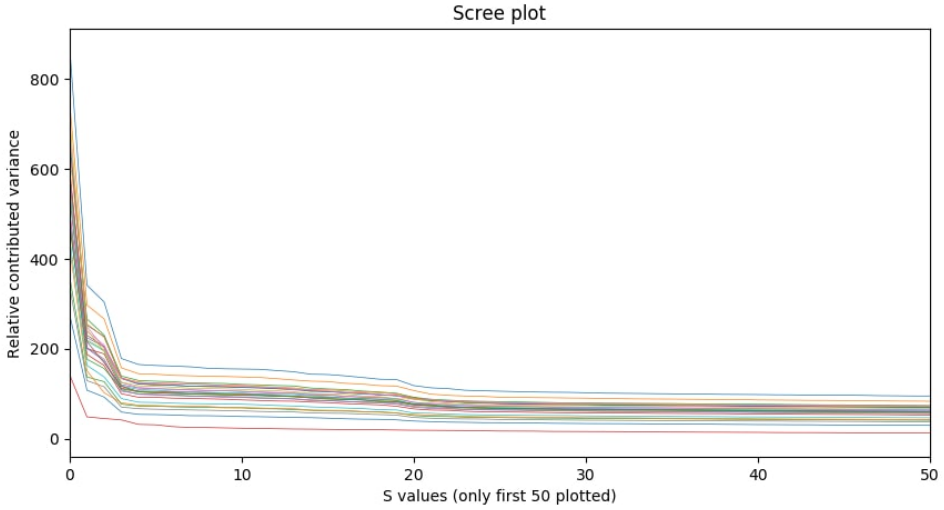


FIGURE 3.7: Scree plot of the singular values. The inflection occurs at 6 and these components were removed.

3.2.3 Quality Control

The depth-normalization approach in CoNIFER is similar to XHMM; both effectively use the SVD implementation of PCA to detect and remove large read-depth variations due to non-CNV signals. The methods however significantly diverge when it comes to the normalized data. CoNIFER makes calls on the basis of consecutive runs of at least three targets with values above or below a hard threshold, whereas XHMM uses HMMs to make and assess the quality of the CNV calls. As per the default settings a SVD value of 6 was used to remove the six principal components, and visual inspection of the singular values in fig 3.7 confirms that this was near the inflection point of the scree plots.

Chapter 4

Results & Analysis

4.1 XHMM

4.1.1 Distribution of the CNVs called by genome location

The main XHMM output file (.xcnv) gives the details of the sample, type of CNVs called (DEL or DUP), chromosome and interval, size in Kb (kilobases) and other details that are necessary for further analysis. Table 4.1 below outlines the CNVs detected in the 94 samples that were ran through the pipeline with default settings, by chromosome, before filtering or genotyping.

TABLE 4.1: Deletions and duplications found in the 94 samples

Variant	Genomic location	Number	Total
DEL	Autosomes	1493	1527
	X	16	
	Y	18	
DUP	Autosomes	1615	1658
	X	30	
	Y	15	

4.1.2 Sensitivity analysis on parameters

In our initial analysis we kept all the XHMM parameter default values to detect the CNVs in the 94 samples. In order to test the robustness of the pipeline and identify critical value ranges for the most crucial parameters in the analysis for our particular

dataset we tested different values for: i) Minimum mapping quality and ii) Maximum standard deviation for the read depth of targets. A lesser and greater value was tested for each parameter, compared to the default value suggested in the pipeline.

4.1.3 Minimum Mapping Quality

This parameter is used in the calculation of depth of coverage per target, the first step of the pipeline. In a probabilistic point of view, each read alignment is an estimate of the true alignment and is therefore also a random variable with an associated error. The error probability scaled in the Phred is the mapping quality. If the mapping quality of a read alignment is mQ, the probability me that the alignment is wrong can be calculated with $me = 10^{-mQ/10.0}$, where me is the mapping error probability.

Several factors are considered for the mapping quality calculation, such as repeat structure of the reference. Indeed, if the reads fall in a repetitive region, the mapping quality will be lower. Low quality implies that the observed sequence could be wrong which may lead to the wrong alignment.

A read alignment with a mapping quality of 30 implies the following:

- The overall base quality of the read is good.
- The best alignment has few mismatches.
- The read has few or just one good hit on the reference, which means the current alignment is still the best, even if one or two bases are actually mutations or sequencing errors.

The default value for the minimum mapping quality suggested by the XHMM pipeline is 20. We tested values of minimum mapping quality of 15 and 25 and compared the results.

After running GATK to detect depth of coverage for each target we get four output files:

- _summary: total, mean, median, quartiles, and threshold proportions of read depth, aggregated over all bases

- _statistics: coverage histograms (# locus with X coverage), aggregated over all bases
- _interval_summary: total, mean, median, quartiles, and threshold proportions of read depth, aggregated per interval
- _interval_statistics: 2x2 table of # of intervals covered to X depth in Y samples

We extracted the granular median of read depth for each of the 10 groups with the three different mapping quality thresholds for the 94 samples from the files with summary and interval summary. Only 5 samples out of 94 showed some difference in the results obtained with different values of minimum mapping quality (Table 6, Appendix) and thus changing this parameter within the tested range did not have a serious impact in the results.

To further check the results, we compared the average coverage for each sample, split by chromosome to make the analysis more agile given the size of the files. We focused on the Y chromosome as that is of major interest to us given the large number of genes important for male fertility but once again there was no significant difference for the three mapping qualities tested.

The graph in fig 4.1 depicts the results and it is evident that the results obtained with the three different values of minimum mapping quality are almost indistinguishable.

4.1.4 Maximum Standard Deviation of the target read depth

A change in the standard deviation for the target read depth. After the PCA normalization, there were still a number of targets with extreme variability in normalized read depth. By default, XHMM filters out targets with a standard deviation of normalized read depth > 30 to remove any outliers from the analysis. The factors that most contribute to variance in target read depth are exome capture protocols and thus different experiments may need more or less stringent filtering of outliers. Thus this parameter may need to be adjusted in our experiment.

As a way of testing the impact of using different thresholds for the standard deviation for target read depth we compared the number of CNVs called with each sd value tested

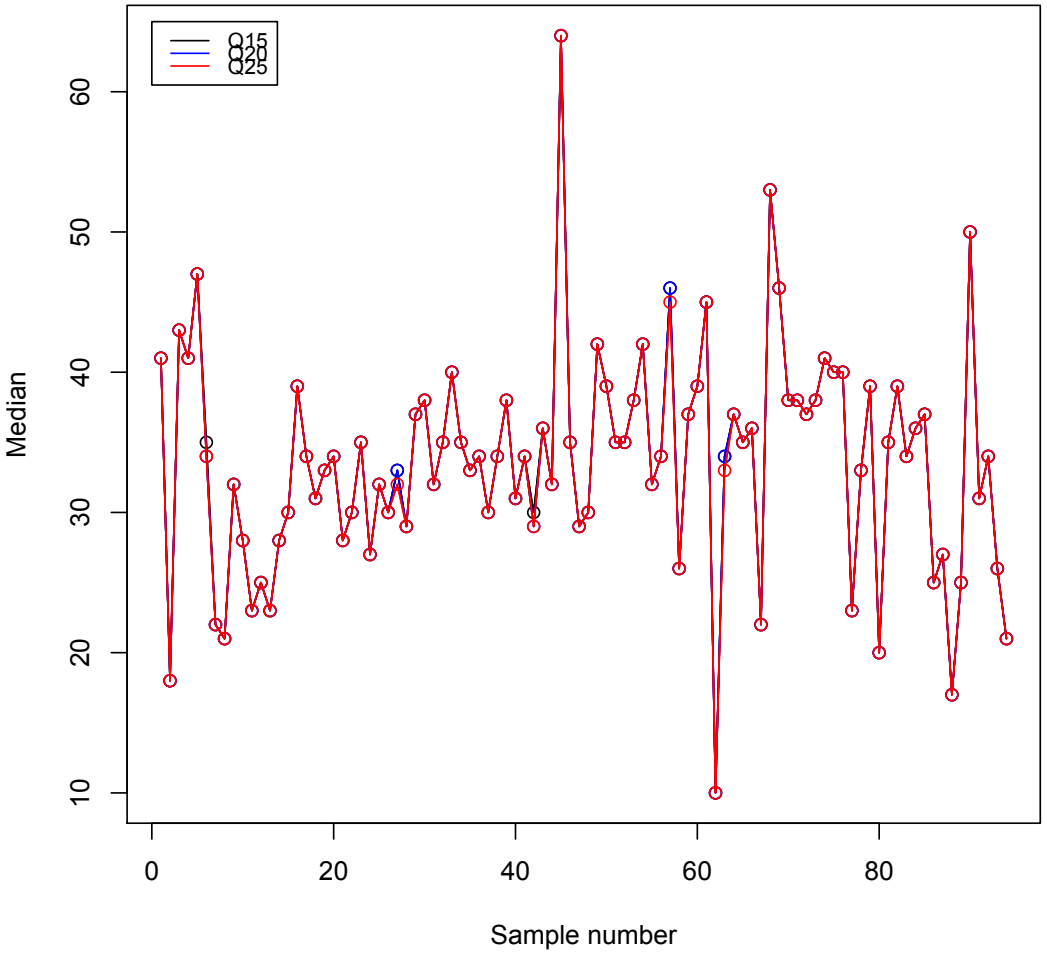


FIGURE 4.1: Graph for the three granular medians for each sample for the three different mapping quality values (Q15 black, Q20-blue, Q25-red). Total of 94 samples

(20, 30 and 40). Our expectation was that a more stringent cutoff for sd would result in less CNVs being called because more targets would be removed from the analysis. On the other hand, if we increased the cutoff we would keep in the analysis targets that show more variable read depth values and likely would include those regions of the genome which are more prone to experimental biases, such as repetitive regions. However, some of these regions may be of interest to our study, even though the results have to be interpreted with caution.

Tables 4.2 and 4.3 below outlines the detection of the deletions and duplication's for the three different read depth standard deviation thresholds.

TABLE 4.2: Table showing the duplication for the three different standard deviation thresholds for target read depths

Chromosomes	**Sd<20**	**Sd<30**	**Sd<40**
X	6	16	19
Y	2	18	41
Autosomes	650	1493	2010
Total	658	1527	2070

TABLE 4.3: Table showing the duplication for the three different standard deviation thresholds for target read depths

Chromosomes	**Sd<20**	**Sd<30**	**Sd<40**
X	24	30	41
Y	7	13	14
Autosomes	764	1615	2081
Total	795	1658	2136

It can be seen that, the more stringent we were with the standard deviation allowed in the CNV calling, the less CNVs were detected.

Finally, we wanted to determine the characteristics of the CNVs that were detected when we were more permissive in this step of target filtering. Since one of our main focus is the Y chromosome, we started by enquiring if the additional CNVs being called in this chromosome were located within repetitive regions, which are more prone to variation in read depth. Since we did not filter out the low complexity regions in the earlier steps of the analysis, as suggested in the pipeline, it was pertinent to test this now.

We used betools to intersect the .xcnv files with other files with genomic regions annotated by feature (such as Y chromosome repeats), which were downloaded from the Table Browser at the UCSC Genome Browser (https://genome.ucsc.edu/). To tests if there was an over-representation of Y repeats between calling sets obtained with different standard deviation thresholds, Fishers exact test was conducted. Intersection of Y chromosome repeats with the .xcnv files for the three different standard deviation values individually with a Fishers exact test conducted in all the three instances was not significant for any ($p>0.05$).

The sequential intersection of each xcnv file obtained with different standard deviation values (sd>20 with sd>30; sd>30 with sd>40) resulted in output files containing a list of regions corresponding to several of the multicopy genes on the human Y chromosome (e.g RPS4Y1,ZFY,AMELY). This indicates that the CNVs being called correspond to

Y genes that are present in more than one copy in healthy individuals and since XHMM does not give exact copy number, but only detects duplications and deletions, these results have to be taken with care.

Next we tested whether the Y chromosome CNVs called using different standard deviation values are located within targets with lower/higher read depths. Therefore from the files (PT.DATA.xcnv (sd20, 30,40)) we filtered the Y chromosome targets and took the columns which had mean_RD(normalized) and mean_original_RD and calculated the average for each. The table below(4.4) summarizes the result:

TABLE 4.4: Mean read depths for Y chromosome targets before and after normalization

sd	**Normalized_Mean read depth**	**Original Mean read depth**	**Total CNV**
<20	3.546	51.746	9
<30	-0.217	56.728	31
<40	-2.434	40.41	55

From the table it can bee seen that for the normalized mean, as the standard deviation increases the read depth decreases which is not the case for the original mean read depth. The original mean read depth increases slightly for the standard deviation of 30, however it decreases as we increase the maximum standard deviation to 40.

To test if this fluctuation between analysis is statistically significant we conducted an Analysis of Variance (ANOVA) on the normalized and original mean read depth. As suspected the comparison was not significant for the normalized mean read depth however for the original mean read depth the result was significant ($p<0.05$). Since the result for the latter was significant (p=2.94e-07) we further conducted a Tukeys HSD (Honest Significance Difference) test to determine between which two sets does the variation lie. Table 4.5 summarizes the output for the TukeyHSD test and it is apparent that the variation target mean read depth is mainly driven by the CNVs detected defining a maximum standard deviation of 40. This is supported by the fact that upon intersection of the xcnv files with genome targets (IDT), for the CNVs detected using a standard deviation of 40, there are genome intervals corresponding to an additional gene (DAZ4) which does is not present in the sd20 and sd30 analyses.

TABLE 4.5: Tukeys HSD result comparison

Group Variation (Mean original read depth)	P value
Sd30 with sd20	0.0594402
Sd40 with sd20	0.0000009
Sd40 with sd30	0.0011031

4.2 CoNIFER

The main ConIFER output file (.txt) gives the details of the sample, type of CNVs called (DEL or DUP) and chromosome coordinates. Table 1 above outlines the CNVs detected in the 94 samples that were ran through the pipeline with default settings.

Analyzing the results from table 4.6 it can be seen that CoNIFER does not detect any deletion or duplication on the Y chromosome and also the number of CNVs detected is less in count when compared to XHMM. CoNIFER gives a global count of 677 while XHMM gives 3185. Fig 4.2 gives a graphical view of the CNVs detected in length against the frequency. The largest number of CNVs detected were between the range 0-25 kb and most of them were duplications.

TABLE 4.6: Deletions and Duplications found in 94 individuals using CoNIFER

Variant	Genomic location	Number	Total
DEL	Autosomes	93	98
	X	5	
	Y	0	
DUP	Autosomes	531	579
	X	48	
	Y	0	

In order to find the rare CNVs a basic CoNIFER uses threshold defining algorithm on the genomic coordinates from the calls listed in output file (calls.txt). The discovery thresholds are set at -1.5 or +1.5 SVD-ZRPKM[12] for rare deletions and duplications, respectively, and required at least three exome probes to exceed the threshold out of the 205574 probes.

Fig 4.3, 4.4 shows a graphical representation of a deletion and duplication call made using the SVD-ZRPKM data. The following point gives an explanation of the plot:

- X-axis: Exons/probes from the probes.txt file
- Y-axis: SVD-ZRPKM values for each exon

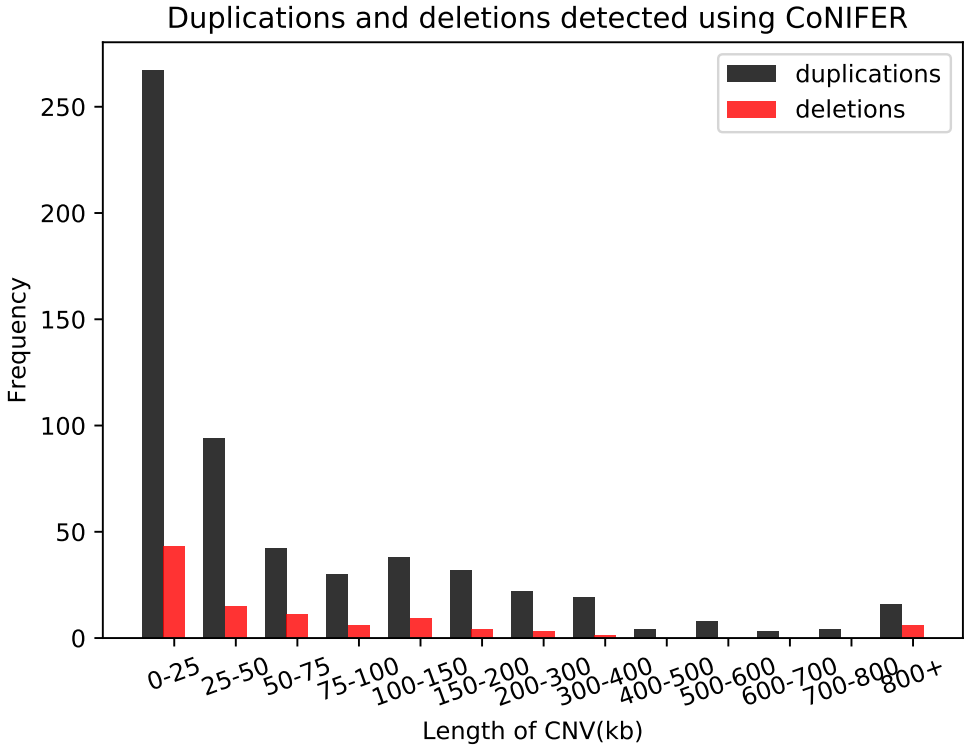


FIGURE 4.2: CNVs detected using CoNIFER

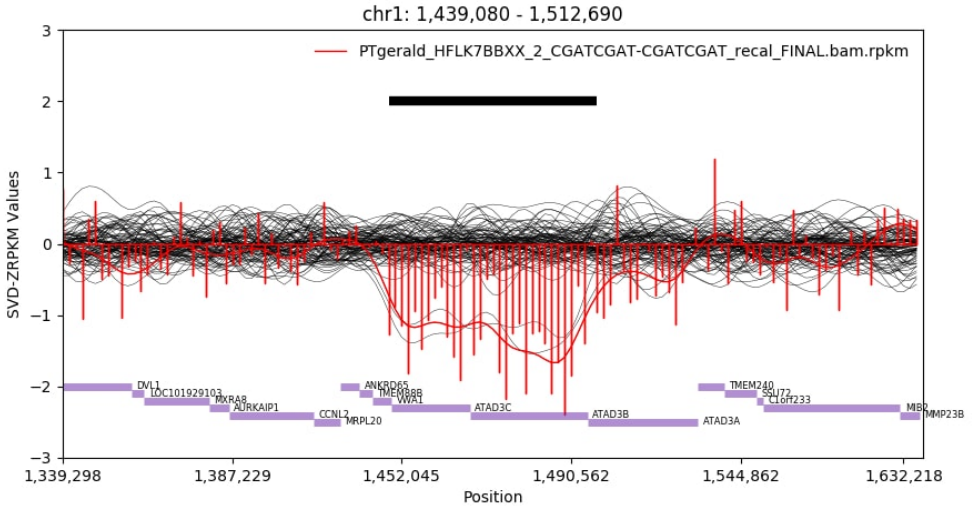


FIGURE 4.3: Deletion call made on chromosome 1; for more details, see main text

- Red bars and line: SVD-ZRPKM values for each exon from the sample with the call. The smooth continuous line is a gaussian-smoothed representation of the SVD-ZRPKM values at each exon

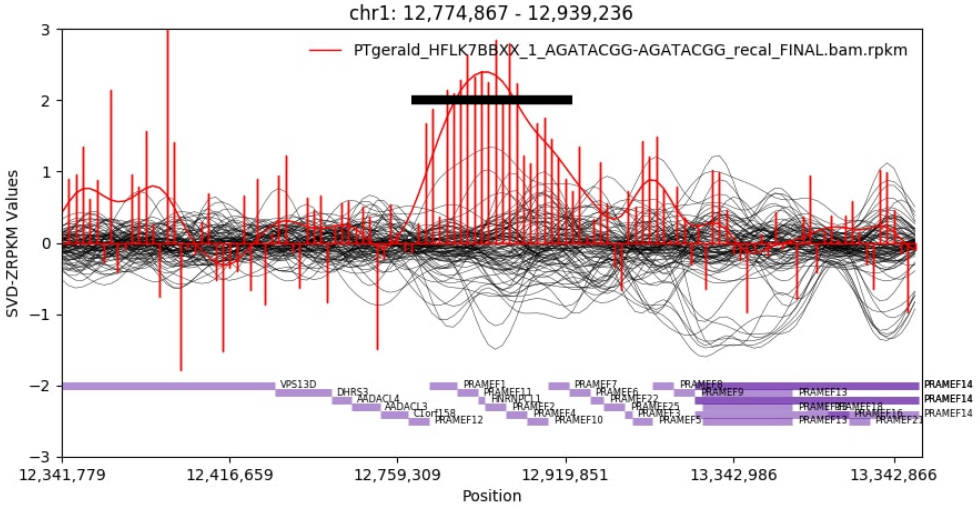


FIGURE 4.4: Duplication call made on chromosome 1; for more details, see main text

- Thin black in background: All other samples in the analysis (for comparison and simultaneous visualization purposes)
- Purple bars: Genes from the probes.txt file
- Black bar: this bar indicates extent of call above threshold (threshold) value

4.3 Comparison of Xhmm and Conifer

There are a number of bioinformatics programs that are able to call CNVs from exome or whole genome sequencing data, providing a fast and cost effective method to detect CNVs. Copy number variants (CNVs) are one of the genetic changes with more potential for a large effect on the phenotype and play a big role in the development of individual traits and in disease susceptibility. Therefore it is of paramount importance to apply an efficient tool to make the calls as accurate as possible. For this thesis two pipelines were tested: XHMM and CoNIFER which incorporate algorithms that differ in the parameters used for assessing coverage in a particular region (table 4.7. But both algorithms take the coverage of a particular area in all given samples and from that data derive whether the area is a CNV in any of the samples or not. XHMM uses the average coverage of each exon for calculations and CoNIFER counts the number of reads in a given exon, however both these tools use a similar procedure consisting of two stages: a normalization step to mitigate systematic biases due to GC content, mappability and capture efficiency, and a segmentation step for the identification of the boundaries of the altered region and the estimation of the absolute or relative number of DNA copies [13]. CoNIFER and XHMM employ singular value decomposition (SVD) and principal-component analysis (PCA) techniques, respectively, to identify and remove the major sources of variation underlying the non-uniform read depth among captured regions. SVD and PCA normalization procedures require the analysis of many samples at once, which can be challenging for a large-scale sequencing project in terms of the memory capacity and the processing time.

Table 4.7: Summary of methods used by XHMM and CoNIFER

CNV caller	Pre-processing quality control	Approach to discovering CNVs
XHMM	Filter extreme GC content (<0.1 or >0.9), lowcomplexity (>10 %), target size (<10 kb or >10), samples (mean RD 30),targets (Mean RD 30). SVD-PCA normalization, remove K components = 0.7/n s	Z-score calculation as input for three-state HMM
CoNIFER	RPKM for each target (filter targets with median RPKM <1), ZRPKM, SVD-PCA transformation. Filter samples >0.5 SVD-ZRPKM	1.5 SVD-ZRPKM threshold values

In this section we compare the results obtained using the two different algorithms. Table 4.8 shows the total number of deletions and duplications detected using the two callers.

TABLE 4.8: Total number of deletions and duplication's called by each platform with default parameters

Variant	**Genomic location**	**XHMM**	**CoNIFER**
Deletions	Auto	1493	93
	X	16	5
	Y	18	0
	Total	**1527**	**98**
Duplications	Auto	1615	531
	X	30	48
	Y	13	0
	Total	**1658**	**579**

For the 94 samples CoNIFER failed to make any calls on the Y chromosome whereas Xhmm made 31 calls. Fig 4.5 shows the graphical comparisons for the calls made using the two algorithms.

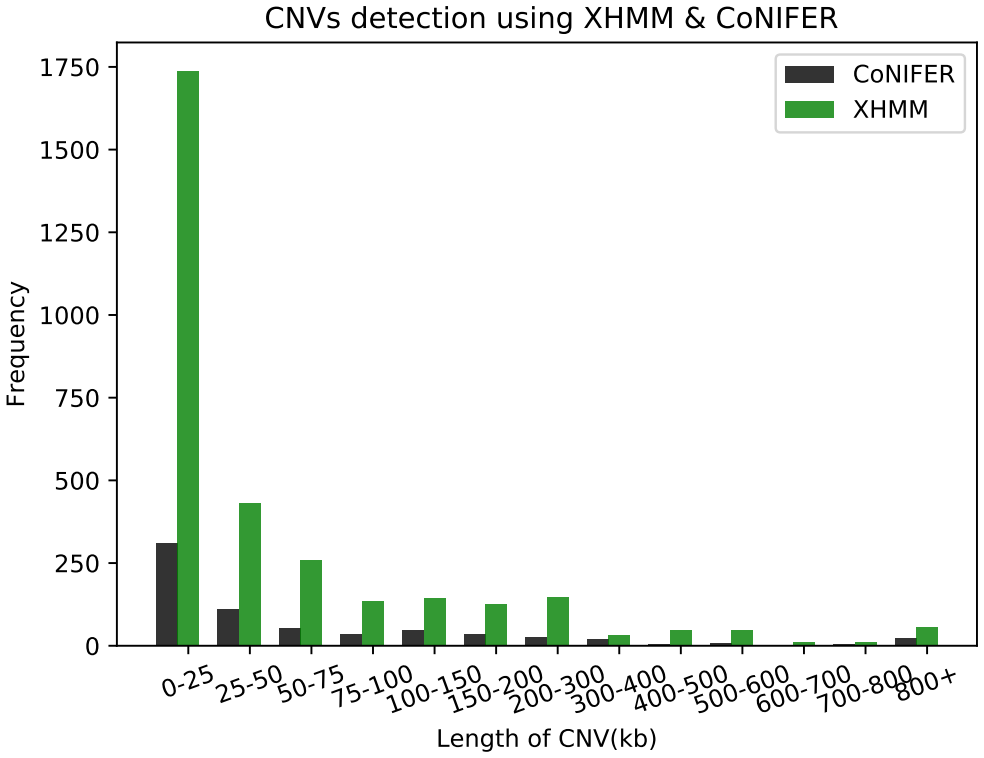


FIGURE 4.5: CNVs detected using CoNIFER;black bars represents detection made by CoNIFER and green represent detection made by XHMM

Comparing just the total number of variants detected by the two callers was not adequate to gauge the two callers performance ability. For example CoNIFER detected

approximately 21.3% of the call counts made by XHMM however based on the call count alone, we could not make comparisons. Some other factors which were of interest; whether there are any similarities in the calls made by the two algorithms, do they fall in the same region, and how reliable are these calls when compared to the CNVs that have been called using whole genome data from the 1000 Genomes[14].

The focus here is to identify/detect the rare variants. CoNIFER characterizes the two distinct classes of genetic variation as rare if the frequency is less than or equal to 1% while Xhmm does the same with a frequency of less than or equal to 5% which means CoNIFER is more stringent.

In order to find the relationship between these two genome-wide datasets we used Bedtools.

4.3.1 Results

First we measured the central tendency for each caller. XHMM produced 3185 calls with mean of 123.86 kb and median of 18.59 kb while CoNIFER produced 677 calls with mean of 175 kb and median of 28.76 kb.The spread was 865.68 kb and 1143.61 kb respectively. This points to the fact that the calls made by the two callers are highly dispersed. The coefficient of variation for both was greater than one. XHMM had 48% deletions and 52% duplications while CoNIFER had 14.5% deletions and 85.5% duplications.

As mentioned earlier in the section, the output file for xhmm was in .xcnv format and for conifer it was in .txt format. Bedtools only reads files in .bed format therefore both output files had to be converted accordingly.

In order to compare the genomic overlaps between the two callers, it was essential to filter the XHMM calls. Using a software tool "Plink", the calls were converted to Plink format and then a filter threshold frequency was set. This was done by finding any CNV that overlaps 50% of another CNV which occurs in more than 10% of all samples[13]. In this way the common CNVs were removed at a threshold of 10%.

From the total calls of 3185 variants, after filtering common CNVs with Plink we obtained 37.5% of the calls, as rare CNVs in our sample with 4 deletions and 10 duplications in the Y chromosome respectively.The filtered XHMM calls had a mean of 182.21 kb and median of 17.88 kb.

XHMM had a higher proportion of calls consisting of between three and five targets (fig 4.5, this is due to ability of the HMM caller to effectively smooth out the normalized read-depth signal and CNVs spanning noisier genomic regions. Since CoNIFER only makes calls when there is atleast 3 more consecutive exons, we filtered the XHMM calls further the only consider the samples which spanned at least 3 exons. This was then used to compare overlaps and intersection with calls made from CoNIFER. We considered 50% overlap of CoNIFER calls over XHMM and got 328 hits.

Next, we took advantage of the well characterized sample NA12878 where CNVs have been called using whole genome data from the 1000 Genomes and which was included in our exome sequencing and analysis, to validate the calls made with both callers. For this sample structural variant (SV) discovery and genotyping was performed on 1000 Genomes Phase 3 using a combination of 9 algorithms designed to identify deletions (DEL), duplications (DUP), multi-allelic copy number variants (mCNV)[14]. Table 4.9 shows the intersection results for CNVs called in NA12878 by the 1000Genomes and duplications and deletions called with XHMM and CoNIFER:

TABLE 4.9: Intersection with cnvs called in NA12878 with 1000G data

Pipeline	**Deletions**	**Duplications**
Xhmm	47	165
CoNIFER	31	156

Lastly we performed the analyses of variance test (ANOVA) for the two datasets; filtered and unfiltered. This was done to compare the mean length of the CNV detected by both the callers. The results were non-significant ($p>0.05$) for both which means we do not have enough evidence to conclude that the means are equal.

Chapter 5

Conclusion and future work

5.1 Conclusion

In this thesis work two pipelines were tested; XHMM and CoNIFER to normalize the coverage in exome sequencing and discover CNVs.

As the first caller we used XHMM to detect CNVs from exome sequence data with a focus detection of under 5%. This was performed to detect rare variants in our cohort of infertile men. XHMM uses principal component analysis (PCA) to normalize the read depth and Hidden Markov model (HMM) to detect the CNV. It categorizes the segment of exome in diploid, deletion or duplication regions, which corresponds to average, below average and above average read depths. We used 94 samples with default settings and detected 31 CNVs on the Y chromosome. To explore the sensitivity of the mapping quality we changed the default parameter and repeated the process. There was no major difference in the results. For the detection of CNV, the read depths need to be homogenous. After PCA is performed, there were still some targets that had high variability. By default XHMM filters out targets with a standard deviation of the normalized read depth grater than 30. We changed the parameters to see the difference in the detection of CNV and concluded that the more stringent we get with the filtering, the less CNVs are detected.

In comparison CoNIFER detects has a focus detection of less that 1% for rare CVNs and uses z-transformation paired with singular value decomposition to make calls. The nature of z-transformation paired with SVD makes this algorithm unsuitable for the

detection of chromosomal aneuploidy as it processes each chromosome separately, and extremely large events are likely to be normalized as part of the first few components.[12]

Comparing the results more calls were made in the interval of 0-25 kb using XHMM as it uses HMM for smoothing over noisy regions and taking into account exome-wide CNV rates.

By examining the exome CNV predictions it can be seen from the results that the reproducibility of the two CNV caller predictions was poor. While CNV callers are potentially valuable and play a major role in identifying CNVs that may be associated with disease, it can be concluded from the results produced in this thesis that there is still a need for improvement.

Another factor to be considered is that these softwares rely on a large number of samples. One reason for the results not being consistent could be due to the fact that we analyzed 94 samples. A fixed threshold cutoff may not be appropriate for all sample sizes due to increase in the noise as more data are added. An adjustment for increased sample size may be appropriate for a large population study[15]. After running 500 samples we expect more consistent results between the callers.

5.2 Future work

As this is an ongoing project and there is still a huge subset of samples to be analyzed, with time we expect to cover the following:

- Test more pipelines
- Analyze the remaining samples
- Statistically assess discovered CNVs in all samples (Genotyping)
- Perform association tests with case and control data and determine if CNVs in coding regions of the exome are a risk factor for azoospermia

For the next pipleline we will use SNP genotypes and read depth to call deletions in the samples. Diploid organisms have the same loci on each of their two sets of homologous chromosomes except that the sequences at these loci may differ between the two chromosomes in a matching pair and that a few chromosomes may be mismatched as part of a chromosomal sex-determination system. If both alleles of a diploid organism are the same, the organism is homozygous at that locus. If they are different, the organism is heterozygous at that locus. If one allele is missing, it is hemizygous, and, if both alleles are missing, it is nullizygous.

Homozygous and hemizygous deletion Finder (HMZDel Finder) extracts read depth from bam and pre-proceeded vcf files to identify regions of deletions.

We aim to explore as many tools as possible which can assist in the precise calling of the CNV in azospermic patients and then compare these with the controls (healthy male samples). Based on the results we then can conclude whether CNV variations contribute to infertility in males.

Appendix A

Appendix

TABLE A.1: Granular median comparison for the three different mapping quality. Table shows the 5 medians that did not match

sample_id	**q15**	**granular_median_re-run q20**	**q25**	**status**
H_VZ-E17-E17	35	34	34	nomatch
mH_VZ-Y1728_05-Y1728_05	33	33	32	nomatch
H_VZ-Y1819_05-Y1819_05	30	29	29	nomatch
H_VZ-Y2425_07-Y2425_07	46	46	45	nomatch
H_VZ-Y1825_05-Y1825_05	34	34	33	nomatch

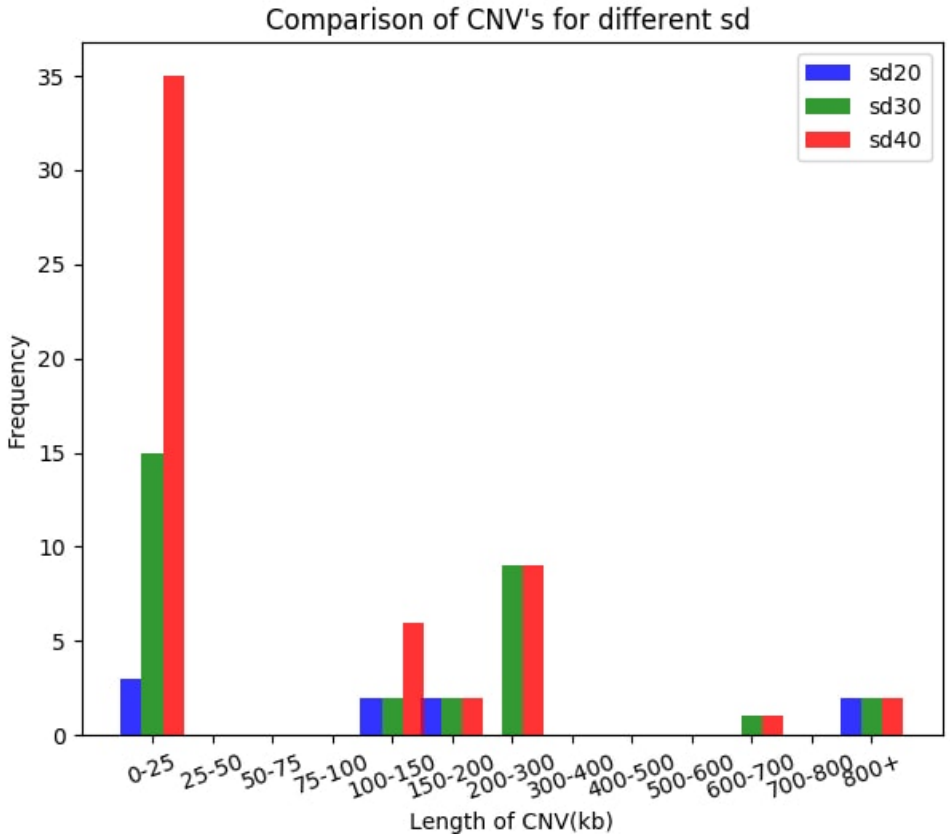


FIGURE A.1: Graph for the number of detected CNVs on Y chromosome using different sd; blue bars represents the detection made using sd20, green bars represents the detection made using sd30 and red bars represents the detection made using sd40

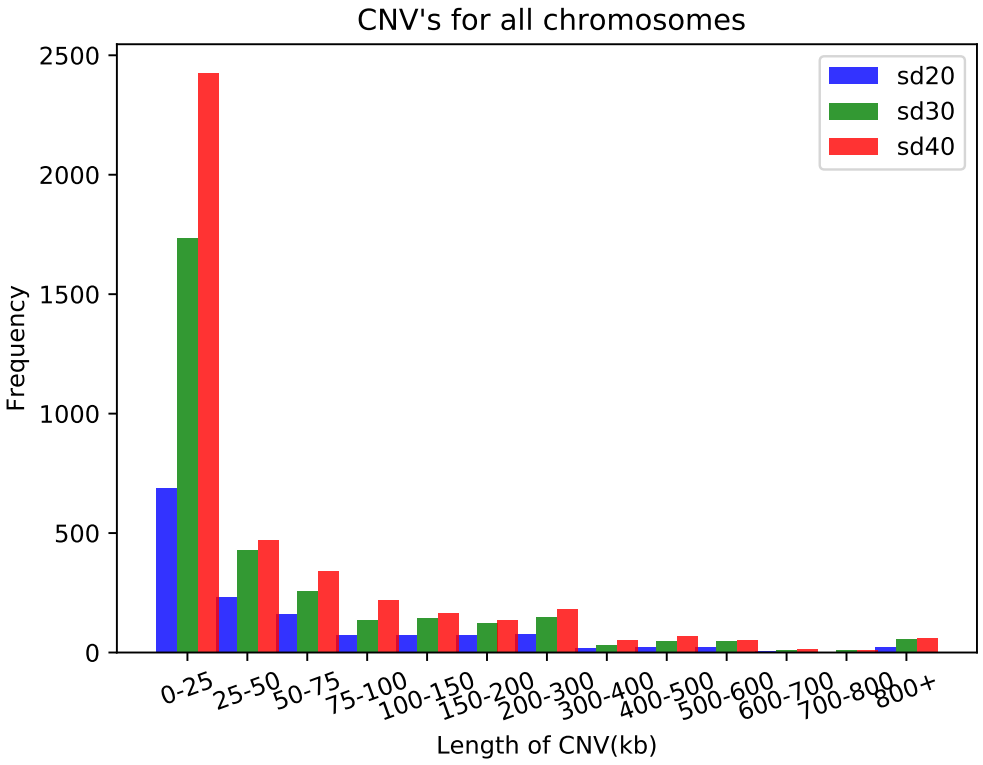


FIGURE A.2: Graph for the number of detected CNVs on all the chromosomes using different sd; blue bars represents the detection made using sd20, green bars represents the detection made using sd30 and red bars represents the detection made using sd40

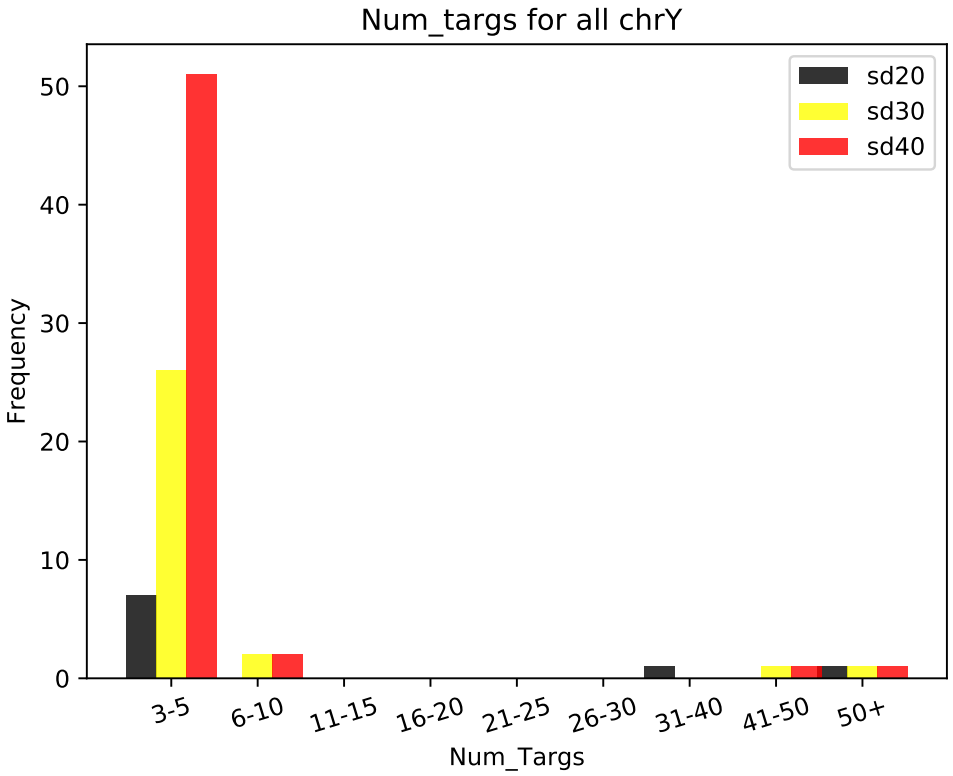


FIGURE A.3: Graph for the target captured(number of exons) in Y chromosome using different sd; blue bars represents the detection made using sd20, green bars represents the detection made using sd30 and red bars represents the detection made using sd40

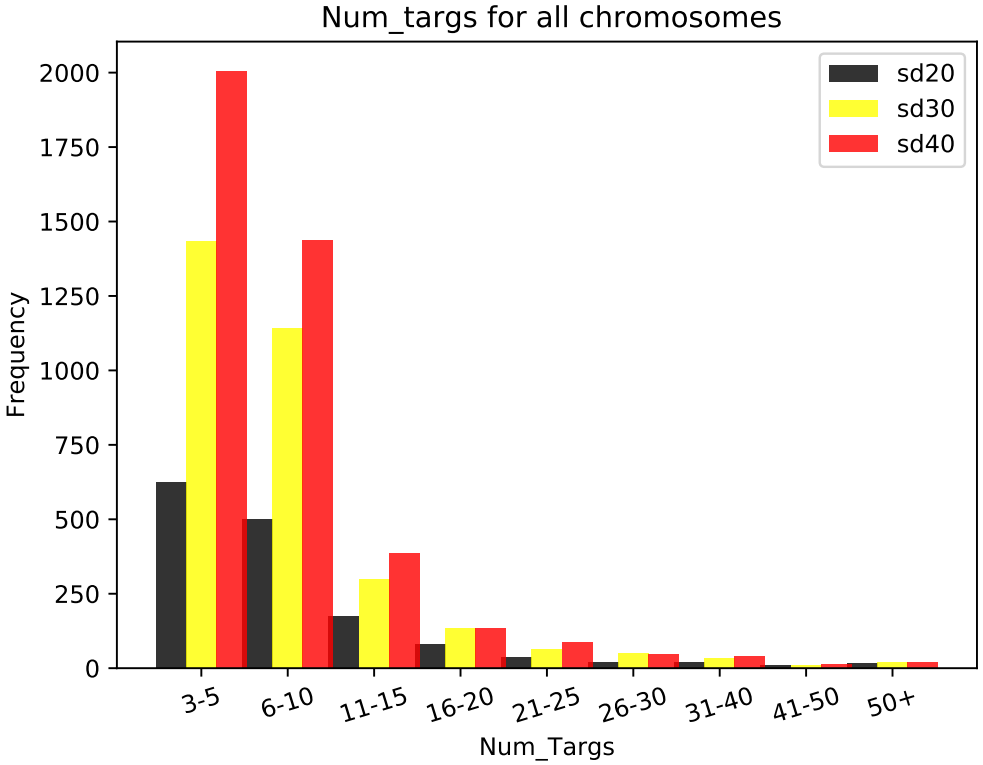


FIGURE A.4: Graph for the target captured(number of exons) on all the chromosomes using different sd; blue bars represents the detection made using sd20, green bars represents the detection made using sd30 and red bars represents the detection made using sd40

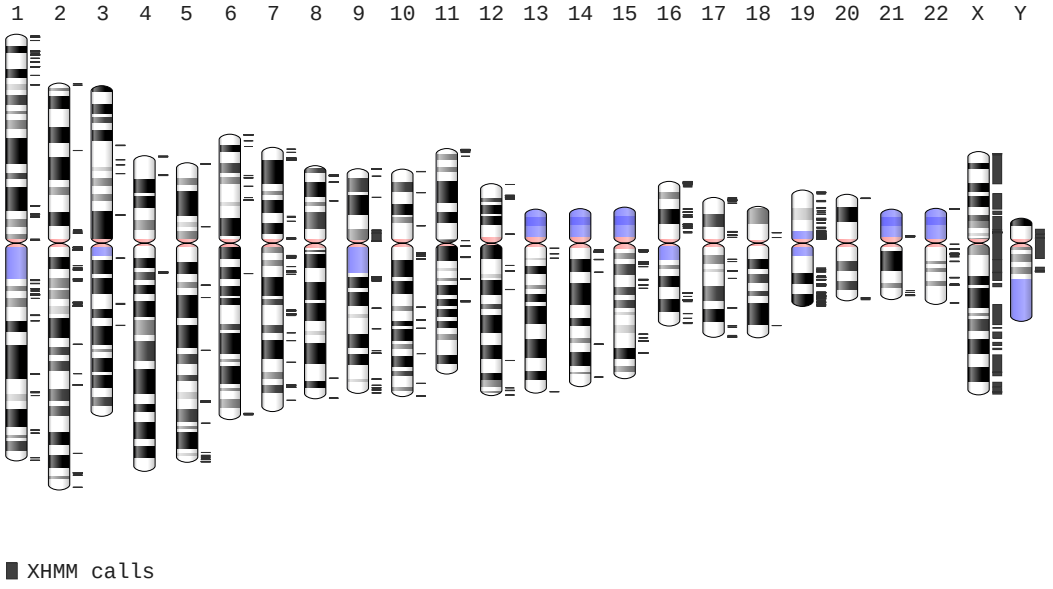


FIGURE A.5: Call distribution across the genome made by XHMM

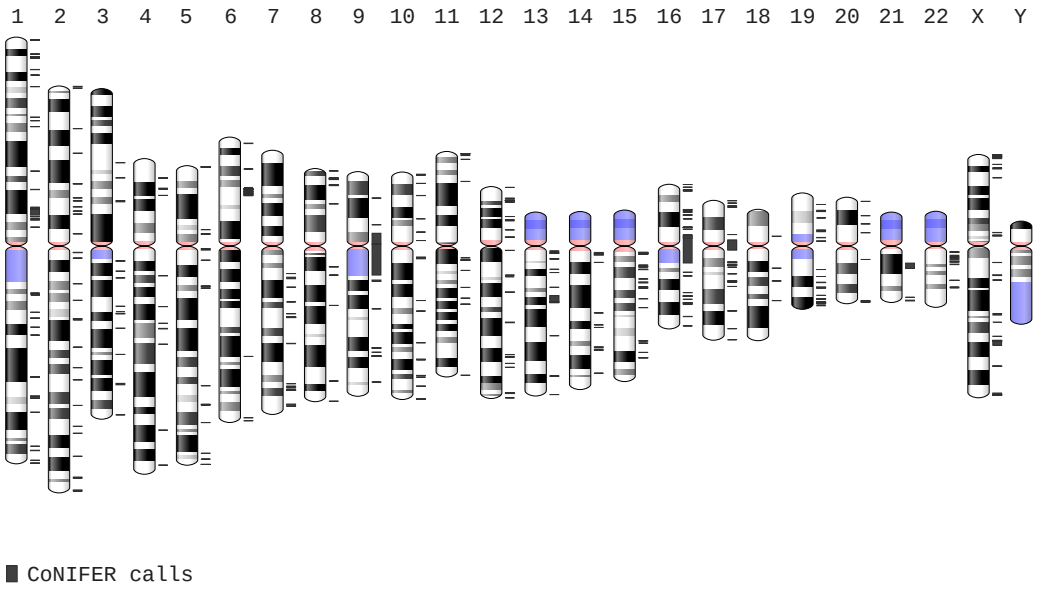


FIGURE A.6: Call distribution across the genome made by CoNIFER

Bibliography

[1] M.A. Jobling, M. Hurles, and C. Tyler-Smith. *Human Evolutionary Genetics: Origins, Peoples & Disease*. Garland Science, 2004. ISBN 9780815341857. URL https://books.google.pt/books?id=104VAAAAIAAJ.

[2] Kenneth I. Aston, Csilla Krausz, Ilaria Laface, E. Ruiz-Castan, and Douglas T. Carrell. Evaluation of 172 candidate polymorphisms for association with oligozoospermia or azoospermia in a large cohort of men of european descent. *Human Reproduction*, 25(6):1383, 2010. doi: 10.1093/humrep/deq081. URL +http://dx.doi.org/10.1093/humrep/deq081.

[3] Ashok Agarwal, Aditi Mulgund, Alaa Hamada, and Michelle Renee Chyatte. A unique view on male infertility around the globe. *Reprod Biol Endocrinol*, 13: 37, Apr 2015. ISSN 1477-7827. doi: 10.1186/s12958-015-0032-1. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4424520/. 32[PII].

[4] Alexandra M. Lopes, Kenneth I. Aston, Emma Thompson, Filipa Carvalho, Joo Gonalves, Ni Huang, Rune Matthiesen, Michiel J. Noordam, Ins Quintela, Avinash Ramu, Catarina Seabra, Amy B. Wilfert, Juncheng Dai, Jonathan M. Downie, Susana Fernandes, Xuejiang Guo, Jiahao Sha, Antnio Amorim, Alberto Barros, Angel Carracedo, Zhibin Hu, Matthew E. Hurles, Sergey Moskovtsev, Carole Ober, Darius A. Paduch, Joshua D. Schiffman, Peter N. Schlegel, Mrio Sousa, Douglas T. Carrell, and Donald F. Conrad. Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene dmrt1. *PLOS Genetics*, 9(3):1–16, 03 2013. doi: 10.1371/journal.pgen.1003349. URL https://doi.org/10.1371/journal.pgen.1003349.

[5] Chiara Chianese, Adam C. Gunning, Claudia Giachini, Fabrice Daguin, Giancarlo Balercia, Elisabet Ars, Deborah Lo Giacco, Eduard Ruiz-Casta, Gianni Forti, Csilla

Krausz, and Eduard Ruiz-Castañé. X chromosome-linked cnvs in male infertility: Discovery of overall duplication load and recurrent, patient-specific gains with potential clinical relevance. *PLoS One*, 9(6):e97746, Jun 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0097746. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4051606/. PONE-D-13-51133[PII].

[6] Ni Huang, Yang Wen, Xuejiang Guo, Zheng Li, Juncheng Dai, Bixian Ni, Jun Yu, Yuan Lin, Wen Zhou, Bing Yao, Yue Jiang, Jiahao Sha, Donald F. Conrad, and Zhibin Hu. A screen for genomic disorders of infertility identifies mast2 duplications associated with nonobstructive azoospermia in humans. *Biol Reprod*, 93(3):61, Sep 2015. ISSN 0006-3363. doi: 10.1095/biolreprod.115.131185. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4710186/. Manuscript Number: BIOLREPROD/2015/131185[PII].

[7] Wenli Li and Michael Olivier. Current analysis platforms and methods for detecting copy number variation. *Physiol Genomics*, 45(1):1–16, Jan 2013. ISSN 1094-8341. doi: 10.1152/physiolgenomics.00082.2012. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3544484/. PG-00082-2012[PII].

[8] P. Dharanipragada and N. Parekh. Copy number variation detection workflow using next generation sequencing data. In *2016 International Conference on Bioinformatics and Systems Biology (BSB)*, pages 1–5, March 2016. doi: 10.1109/BSB.2016.7552117.

[9] Nicholas J Schork, Sarah S Murray, Kelly A Frazer, and Eric J Topol. Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics Development*, 19(3):212 – 219, 2009. ISSN 0959-437X. doi: https://doi.org/10.1016/j.gde.2009.04.010. URL http://www.sciencedirect.com/science/article/pii/S0959437X09000884. Genetics of disease.

[10] Min Zhao, Qingguo Wang, Quan Wang, Peilin Jia, and Zhongming Zhao. Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14(Suppl 11): S1–S1, Sep 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-S11-S1. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3846878/. 1471-2105-14-S11-S1[PII].

[11] Menachem Fromer and Shaun M. Purcell. Using xhmm software to detect copy number variation in whole-exome sequencing data. *Curr Protoc Hum Genet*, 81:7.23.1–7.23.21, Apr 2014. ISSN 1934-8266. doi: 10.1002/0471142905.hg0723s81. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4065038/. 24763994[pmid].

[12] Niklas Krumm, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, NHLBI Exome Sequencing Project, Aaron R. Quinlan, Deborah A. Nickerson, and Evan E. Eichler. Copy number variation detection and genotyping from exome sequence data. *Genome Res*, 22(8):1525–1532, Aug 2012. ISSN 1088-9051. doi: 10.1101/gr.138115.112. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409265/. 22585873[pmid].

[13] Menachem Fromer, Jennifer L Moran, Kimberly Chambert, Eric Banks, Sarah E Bergen, Douglas M Ruderfer, Robert E Handsaker, Steven A McCarroll, Michael C ODonovan, Michael J Owen, George Kirov, Patrick F Sullivan, Christina M Hultman, Pamela Sklar, and Shaun M Purcell. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet*, 91(4):597–607, Oct 2012. ISSN 0002-9297. doi: 10.1016/j.ajhg.2012.08.005. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3484655/. AJHG1232[PII].

[14] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct 2015. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature15393. Article.

[15] Celine S. Hong, Larry N. Singh, James C. Mullikin, and Leslie G. Biesecker. Assessing the reproducibility of exome copy number variations predictions. *Genome Medicine*, 8(1):82, 2016. ISSN 1756-994X. doi: 10.1186/s13073-016-0336-6. URL http://dx.doi.org/10.1186/s13073-016-0336-6.